Semantic component selection.

Maxym Sjachyn

School of Electronics and Computer Science

SEMANTIC COMPONENT SELECTION

MAXYM SJACHYN

A thesis submitted in partial fulfilment of the
Requirements of the University of Westminster
for the degree of Doctor of Philosophy

August 2009

Declaration

I hereby declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly while in candidature for a research degree at University of Westminster;
- This thesis has not been previously submitted for a degree or any other qualification at this University or any other institution;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;

Parts of this work have been published as:

Sjachyn, M. and L. Beus-Dukic (2006). Semantic component selection - SemaCS. 5th IEEE International Conference on COTS-Based Software Systems, Orlando, Florida, USA, IEEE Computer Society Press.

Acknowledgements

It is commonly said that writing a PhD dissertation is a lonely process. However, as with any journey, this does not have to be the case. Consequently my sincere gratitude goes:

To Natalie Sjachyna, my mother, for putting up with 4 years of continual occupation of all IT resources;

To Ljerka Beus-Dukic, my DOS, for always being there to help, encourage and guide me;

To Simon Courtenage, 2nd supervisor, for his continual help and support;

To Andrzej Tarczynski, 3rd supervisor, for his support, help and encouragement;

To Marko Dukic, for proofreading my thesis;

To the university and all the staff, students, friends and Sensei Will Wimshurst and Simon Ly for making this journey not a lonely one.

It is the journey and the people travelling alongside that define us.

Abstract

The means of locating information quickly and efficiently is a growing area of research. However the real challenge is not related to locating bits of information, but finding those that are relevant. Relevant information resides within unstructured 'natural' text. However, understanding natural text and judging information relevancy is a challenge. The challenge is partially addressed by use of semantic models and reasoning approaches that allow categorisation and (within limited fashion) provide understanding of this information. Nevertheless, many such methods are dependent on expert input and, consequently, are expensive to produce and do not scale. Although automated solutions exist, thus far, these have not been able to approach accuracy levels achievable through use of expert input.

This thesis presents SemaCS - a novel nondomain specific automated framework of categorising and searching natural text. SemaCS does not rely on expert input; it is based on actual data being searched and statistical semantic distances between words. These semantic distances are used to perform basic reasoning and semantic query interpretation. The approach was tested through a feasibility study and two case studies. Based on reasoning and analyses of data collected through these studies, it can be concluded that SemaCS provides a domain independent approach of semantic model generation and query interpretation without expert input. Moreover, SemaCS can be further extended to provide a scalable solution applicable to large datasets (i.e. World Wide Web).

This thesis contributes to the current body of knowledge by establishing, adapting, and using novel techniques to define a generic selection/categorisation framework. Implementing the framework outlined in the thesis improves an existing algorithm of semantic distance acquisition. Finally, as a novel approach to the extraction of semantic information is proposed, there exists a positive impact on Information Retrieval domain and, specifically, on Natural Language Processing, word disambiguation and Web/Intranet search.

List of Abbreviations

AI	Artificial Intelligence
ANNIE	a Nearly-New Information Extraction System
CERN	European Council for Nuclear Research
COTS	Commercial-Off-The-Shelf
DC	Data clustering
DM	Data mining
GATE	Generic Architecture for Text Engineering
IDF	Inverse Document Frequency
IR	Information Retrieval
KDD	Knowledge-Discovery in Databases
LE	Language Engineering
LSA	Latent Semantic Analysis
ML	Machine Learning
mNGD	Modified Normalised Google Distance algorithm
MRR	Mean Reciprocal Rank
NGD	Normalised Google Distance algorithm
NL	Natural Language
NLP	Natural Language Processing
OO	Object Orientation
OWL	Web Ontology Language
POS	Parts of Speech
RDF	Resource Description Framework
RF	Relevance Feedback
RR	Reciprocal Rank
SemaCS	Semantic Component Selection
SMART	System for the Mechanical Analysis and Retrieval of Text
SRS	Student Record System
SVM	Support Vector Machine
Taxpet	Taxonomy snippet
TF	Term Frequency
Tier	Depth or granularity level of the model
URL	Uniform Resource Locator
VSM	Vector Space Model
Web	see WWW
WWW	World Wide Web

List of Equations

List of Figures

List of Tables

Table of Contents

Chapter 1 Introduction

Since the first web page was uploaded in 1991, millions of World Wide Web (WWW) users generated billions of web pages. Although the current number of web pages within the WWW can only be estimated (it can even be unlimited if automatically generated pages are included), the size of the indexed WWW can be determined. According to the official Google blog (Google 2008), in 2008, the Google index reached 1 trillion unique URL (Uniform Resource Locator) entries. Such large quantities of available content make the WWW a logical choice for searching; be it for a flight, a car, or scientific publications – it is the first place many users turn to. However, the availability of information which makes the WWW popular also makes it increasingly difficult to search and, as a result, use.

Searching is not a new concept. The need to search has existed long before computers were conceived. When the ability to keep a written record of knowledge appeared, the need to search these records also emerged. Although, physically, computer records and, for example, scrolls used in ancient Greece are inherently different, they are also innately identical. This is because, regardless of the language or the interface used, records are created to store heterogeneous information. Consequently, there is a need to search these records to locate specific information.

1.1 WWW and Information Retrieval

Although the need to search has existed for a long time, this thesis is only concerned with recent computerised developments. As a result, this overview begins with a historical perspective of the WWW, where heterogeneous information is stored, and Information Retrieval (IR), which provides a means to search this information.

1.1.1 A brief history of the WWW

In 1960 Ted Nelson founded project Xanadu (Nelson 1960). Xanadu introduced, in 1963, the concept of, and the term 'hypertext'. Although Xanadu never succeeded, its concepts have, eventually, led to the creation of the WWW.

In 1980, while working at CERN (European Council for Nuclear Research), Tim Berners-Lee (inspired by Xanadu and hypertext concepts) created ENQUIRE – a tool capable of linking information by means of navigable 'hyperlinks' (links of connections between people and projects at CERN). However, ENQUIRE was never made publically available.

In 1989 Tim Berners-Lee proposed a hypertext based Information Management project – a *universal linked information system* (Berners-Lee 1989). Although this proposal was not received well initially, it has, eventually, resulted in the development of the WWW; with the first public web page (http://info.cern.ch/) going 'live' on August the 6th 1991.

While the WWW is of immense importance, it is just a means to store and navigate information. The need to search the information remains and this need is addressed by IR.

1.1.2 A brief history of Information Retrieval

Gerard Salton is considered by many to be a founder of modern search technology (Consroe 1999). In the late 60s he was responsible for the development of the SMART (System for the Mechanical Analysis and Retrieval of Text) IR system (Salton 1971). SMART was pioneering in many respects and

introduced a number of IR methods such as VSM – Vector Space Model (Salton, Wong et al. 1975), IDF – Inverse Document Frequency (Salton, Wong et al. 1975), RF – Relevance Feedback mechanisms (Buckley, Salton et al. 1994) and TF – Term Frequency (Salton 1989). These approaches are still used today and have also, specifically VSM, TF, IDF and RF, served as motivation for the approach to be proposed in this thesis.

With the introduction of the SMART IR system the process of searching has evolved from simple keyword-based algorithms and regular expressions (for example Archie search engine (Emtage and Deutsch 1992)) to something altogether more sophisticated. VSM, although a purely statistical approach, provided the first computerised means of performing computer-based intelligent search.

1.2 Intelligent search and natural text

As with the ancient Greek example from the beginning of this section, given that a number of scrolls containing information of interest are stored in a library, how can a single piece of work by, for example, Aristotle be found? From a human perspective this task can be defined by the following three distinct logical steps:

1. Locate scrolls by Aristotle (section of the library)
2. Identify those of interest (topic of interest)
3. Select those that are relevant

However simple, these 3 steps demonstrate what IR is attempting to achieve and, although such a task may be easy for a person to perform, it has proven difficult to emulate programmatically. There are many reasons why this is a

challenge, but chief amongst them is also one which is the simplest for people to understand – context. Because people can understand document contents, people can reason, infer, and make relevancy decisions – people possess a natural ability to perform an intelligent search.

1.2.1 Computer-based intelligent search

Unlike people, computers do not possess a natural ability to perform an intelligent search. However, at least partially, computer-based intelligent search is possible by, for example, means of: AI - Artificial Intelligence (e.g. (Masterseek 2008)), Ontologies (e.g. (Kim, Alani et al. 2002)), NLP - Natural Language Processing (e.g. (Powerset 2008)), etc. Nevertheless, because a computer-based implementation of intelligent search differs from how it is achieved by a person, the steps taken to achieve it also must differ. According to (Hiemstra 2001), from a computer-based perspective, an IR search task can be split into the following 3 stages:

1. *The representation of the content of the documents*
2. *The representation of the user's information need*
3. *The comparison of the two representations*

Stage 1 can be summarised as the process of indexing, categorising and, possibly, of understanding the document collection; stage 2 can be summarised as the process of adapting, translating or understanding an user query; finally, stage 3 is the process of making relevancy decisions by matching the user query to the relevant documents in the document collection. Each of these stages represents a distinct area of research: for example, Clustering (Maarek, Faginy et al. 2000) or SVMs - Support Vector Machines (Y. Li and Cunningham 2005) can be applied at stage1; query expansion (Xu and Croft 1996) or NLP

(Powerset 2008) can be applied at stage 2; AI (Frank, Hall et al. 2005) or Ontology (Ding, Finin et al. 2005) based reasoning could be applied at stage 3. Although distinct, as a combination, these approaches provide a solution to the problem of searching, understanding and reasoning over natural text.

1.2.2 Natural language and text

Any textual content generated by a person can be considered natural text - it is a written representation of natural language. As natural text follows specific rules (for example this thesis) it could be assumed these rules can be used to understand it. Nevertheless, knowledge of linguistic rules does not necessarily guarantee an ability to use or comprehend a language (or understand a thesis). However, generally with native speakers, linguistic rules may not be known (or are perceived instinctively) but an ability to use the language is displayed. Natural language or its written representation is such a complex entity to understand because it is not defined by a set of simple rules, instead it is a *Complex Adaptive System* (Steels 2007).

Given such complex characteristics of natural text, how can these characteristics be programmatically defined? Through mapping different semantic meanings that words represent? It can be argued that words are used similarly to hieroglyphs, where each hieroglyph denotes some meaning, an understanding of which can only be gained within its complete context (such is the case with 1884 Frege Gottlob (Gottlob 1980) *context principle*). It can also be argued that each word represents one of a possible set of meanings and that its placement or neighbours are irrelevant (this view is generally taken by dictionaries). However, the primary reason of the difficulty of programmatic understanding of natural text is related to it not being grounded in computational logic. Language, both written and spoken, has evolved over thousands of years

by people, not computer programmers; it is a *manifestation of species-specific cognitive propensities* (Lenneberg 1967).

1.3 Scope and objectives

Although many different IR approaches have evolved since keyword-based search techniques (see (Dridi 2008) for an overview), keyword-based algorithms are still commonly used because they offer an efficient, easily implementable solution. However, keyword-based techniques are not capable of understanding the domain being searched. As a result they cannot be used to infer what may or may not be relevant to the user. Consequently, a number of IR approaches designed to find information that are relevant, instead of just finding information, were developed (e.g., NLP driven Powerset (Powerset 2008), AI based Accoona (Masterseek 2008), Clustering based Carrot2 (Stefanowski and Weiss 2003), Ontology based Swoogle (Ding, Finin et al. 2005)). Nevertheless, the need to search is not restricted to the WWW. Increasingly, search of sensitive company documentation, projects, software components, medical information, etc. is sought. Subsequently, many search solutions were developed to provide this functionality (e.g., HP WiseWare (Delic and Hoellmer 2000), IBM OmniFind (IBM 2008) , Vivisimo Velocity (Koshman, Spink et al. 2006)). However, thus far no complete solution has been developed that (see (Qi and Davison 2009) for a detailed comparison) satisfies the following criteria:

- is application domain independent;
- does not require training or rule generation based on expert input;
- does not require an expert generated dictionary or domain taxonomy;
- provides higher accuracy than keyword-based approaches;
- is scalable to large document collections.

Ideas and approaches to be presented in this thesis were originally motivated by the need to provide a solution addressing the above criteria within the software component domain.

1.3.1 Component-based software development

The component-based approach to software development was conceived as an attempt to provide reusability through modularity. In the past, attempts to do so were undertaken by means of C++ Template Libraries and Object Orientation (OO) to name but a few. A component-based approach represents one of the latter attempts to realise the 'promise' of reusability (Szyperski, Gruntz et al. 2002). However, although many Commercial Off-The-Shelf (COTS) software components may be available, they are by definition black boxes so internal structures and implementation are inaccessible. Consequently, suitability decisions are based on textual information provided by the vendor/author and, possibly, other users (Mielnik, Bouthors et al. 2004). As a result, component portals (search platforms) have little information available to provide for categorisation and search. Due to this lack of information, producing representational models and software component categorisation is, usually, a human driven task. Subsequently, the primary aim of this thesis is to define a framework capable of providing a means of automated software component categorisation and selection utilising domain knowledge obtained from textual descriptions of software components being categorised and the WWW.

1.3.2 Objectives

The means of processing and understanding natural text within the software component domain are applicable to many other domains. As a result, this thesis proposes SemaCS (see chapter 3 - SemaCS: Semantic Component Selection) – a framework providing a means of automated categorisation and

search of natural text descriptions (these descriptions can be representative of software components, university modules, web pages, etc.). This objective was represented as primary hypothesis $\mathbf{H_1}$ (see Table 1).

However, to aid with evaluation hypothesis $\mathbf{H_1}$ was further divided into 4 secondary hypotheses (see Table 1). SemaCS is:

- Application domain independent (based purely on free-form textual descriptions and access to the WWW) secondary hypothesis $\mathbf{H_{1a}}$;
- Semantically driven (by means of scale invariant mNGD (2) to acquire semantic distances) secondary hypothesis $\mathbf{H_{1b}}$;
- Capable of automated domain taxonomy generation and search (as a result removing the need for expert input) secondary hypothesis $\mathbf{H_{1c}}$;
- Capable of providing data-based personalisation (consequently improving result relevancy for given task/domain) secondary hypothesis $\mathbf{H_{1d}}$.

H_1 :	**Web-sourced domain knowledge can be applied to automate domain taxonomy generation, categorisation and retrieval processes.**
H_{1a}:	SemaCS approach is not domain specific.
H_{1b}:	mNGD algorithm improves on NGD algorithm by providing a comparable but non N dependant (scale invariant) solution.
H_{1c}:	A representative domain taxonomy can be automatically generated based on mNGD and textual information being categorised.
H_{1d}:	Data-based personalisation has a positive impact on result relevancy.

Table 1: List of hypotheses

These hypotheses were defined (see Chapter 2) as the following research questions:

- Applicability of hierarchical clustering algorithms to large document collections (**H_{1c}**);
- Clustering measure of semantic similarity (**H_{1b}**);
- Classification/understanding of user information need (**H_1** and **H_{1a}**);
- Personalisation as a means to define user information need (**H_{1d}**);

1.3.3 Contribution to knowledge

This thesis is significant for several reasons:

- it defines a novel automated multi-domain natural text search and categorisation framework that does not require expert input to function;
- it improves on (Cilibrasi and Vitanyi 2004; Cilibrasi and Vitanyi 2007) Normalised Google Distance NGD (1) algorithm; modified Normalised Google Distance – mNGD (2) removes NGD (1) dependence on **N** (making it scale invariant);
- it defines a novel method of domain taxonomy generation based on language model approach (Ponte and Croft 1998) and mNGD (2);
- it proposes a novel approach of searching, result personalisation and software component repositories (portals) implementation based on language model approach (Ponte and Croft 1998) and mNGD (2).

This research benefits general users as well as application designers because it defines a novel software component categorisation approach that can provide better matched results then current traditional textual or statistically bound approaches. Furthermore, as a novel approach to extraction of semantic information is proposed, there exists an impact on IR domain and, specifically, on Clustering, word disambiguation, NLP and Web/Intranet search engines.

1.4 Thesis structure

This chapter introduced a component-based software development approach and outlined general problem areas while defining scope, motivation, and the objectives of this thesis. The following chapters expand on concepts introduced here:

Chapter 2 further defines context, motivation, rational and problem domain. This chapter also identifies research questions and describes 3 interrelated areas of research corresponding to the 3 stages of computer-based intelligent search.

Chapter 3 defines a list of inter-related requirements corresponding to hypothesis and research questions identified in chapters 1 and 2. This chapter further describes SemaCS framework and elaborated on design decisions, implementation and algorithms.

Chapter 4 describes chosen methodology rationale and methods used to evaluate primary and secondary hypothesis. This chapter also describes three case studies: Pilot, Sourseforge.net and University of Westminster SRS (Student Record System) module search designed to provide evidence either supporting or negating primary and secondary hypotheses.

Chapter 5 present results collected by means of the three case studies described in chapter 4. This chapter further analyses these results and their significance and impact on primary and secondary hypotheses.

Chapter 6 identifies the main findings of the thesis. This chapter also analyses the impact and significance of mNGD (2) modification and elaborates on evaluation criteria suitability. Additionally, implications for the field and directions for future work are identified.

Chapter 2 Research context

In the previous chapter, motivation, rationale and hypotheses were identified (see Table 1 – List of hypotheses). Chapter 1 has also introduced (section 1.2.1 - Computer-based intelligent search) the 3 stages required to perform computer-based intelligent search. These 3 stages correspond to the 3 interrelated areas of research described here. However, it should be noted that, although some approaches and concepts are easily placed within a specific stage of the intelligent search (for example, in IR taxonomies are generally used to represent a document collection (Prieto-Díaz 1990) or rather the domain of the collection), division of other approaches and concepts is artificial. Consequently, even though this chapter is structured as a representation of the 3 stages of the intelligent search, many approaches and concepts can be applicable to more than one stage (e.g., NLP can be used to understand the contents of the document collection or a user query).

The remainder of this chapter is structured in the following way: section 2.1 introduces concepts and approaches related to stage 1 of the intelligent search: ontologies, taxonomies and data clustering; section 2.2 introduces a number of general concepts and approaches related to stage 2: Data Mining (DM), Machine Learning (ML), NLP and the Semantic web; finally, in section 2.3 concepts and approaches related to stage 3 are described: intelligent search, search engines, software component portals and search result personalisation.

2.1 Stage1: The representation of the content of the documents

This section describes relevant related concepts and approaches corresponding to stage 1 of the intelligent search – representation and understanding of the document collection to facilitate search. Ontological representations are used with search approaches that require knowledge of semantic associations or

knowledge about the domain (e.g., (Dridi 2008), (Graubmann and Roshchin 2006), (Alani, Kim et al. 2003), (Wang, Xu et al. 2004)). Consequently, section 2.1.1 introduces ontologies, taxonomies and a number of related approaches, and further outlines applicability of clustering to stage 1 of the intelligent search. Section 2.1.2 introduces data clustering from the domain representation perspective.

2.1.1 Ontologies

In computer science an ontology is a computer-digestible conceptual model providing the *definition of classes, relations and functions*. An ontology is *a specification of a conceptualization* (Gruber 1993). Ontologies are used in many areas of IR: from ML and DM, to AI, NLP and the Semantic web. However, specifically for IR approaches that search document collections (for example Yahoo! search engine), ontologies provide a *structuring device for an information repository* (Uschold and Gruninger 2004). Arguably, this description is best suited to represent taxonomies, as these are specialised cases of ontologies which provide a hierarchy relation structure that can be used for browsing and classification, but do not provide definitions of objects. Ontologies and taxonomies (specialisation of ontology) can be divided into two broad categories: domain and upper.

2.1.1.1 Domain ontologies

A domain ontology, or domain-specific ontology, is a model of one specific domain (for example CIDOC Conceptual Reference Model ontology (Doerr 2003)), or commonly a part of that domain. Such ontologies tend to contain particular concept definitions as they are used within that domain. For example, the word ‘card’ has many different senses; however, in a games ontology, a card is likely to model a ‘playing card’, while in computer hardware ontology, a card

could represent the concept of a ‘video card’. Domain ontologies have a limited scope of application and are typically created manually. Due to the manually generated nature, they tend to be accurate. However, this manual nature also introduces a severe limitation. Updating or generating new domain ontologies is both expensive and time intensive. As a result, a number of attempts have been made to provide for a level of automation, at least when it comes to reuse. Possibly one of the best known generic examples of such reuse is Swoogle (Ding, Finin et al. 2004).

Swoogle is a crawler-based indexing and retrieval system aimed at the Semantic web community. Unlike general WWW search engines (for example Google) Swoogle indexes RDF (Resource Description Framework) (W3C 2004a) and OWL (Web Ontology Language) (McGuinness and Harmelen 2004) documents. Swoogle extracts metadata from discovered documents and provides ontology rank, a measure of importance based on Google PageRank algorithm (Brin and Page 1998). Swoogle further provides an ontology dictionary compiled from ontologies discovered on the WWW. Similarly to SemaCS (see chapter 3 - SemaCS: Semantic Component Selection) Swoogle extracts knowledge from the WWW but, unlike SemaCS, does not infer knowledge beyond what is described within the found ontologies. It should also be noted that Swoogle has to find a level of intersection between ontologies it locates as well as trust respective sources to provide an accurate definition of their domain. These are the two primary limitations of any approach that attempts to gain a higher level of understanding or descriptiveness by reusing existing ontologies.

Further examples of ontology reuse are provided by (Dridi 2008), (Vogel, Bickel et al. 2005), (Wang, Xu et al. 2004), etc. However, one of the most related to the approach presented in this thesis is a WWW based automated disambiguation method capable of processing *a set of related keywords in order to discover and*

extract their implicit semantics, obtaining their most suitable senses according to their context (Gracia, Trillo et al. 2006). This is achieved by means of ontologies discovered on the WWW, NGD (1) algorithm (see section 3.1.2 – Semantic distance acquisition) and analyses of the sentences. WWW based automated disambiguation method is a novel approach that can be used to model the domain of interest and to understand natural text. However, as it is based on publicly available ontologies, it is subject to the same limitations as Swoogle. Furthermore, NGD (1), in its original form, also introduces limitations (see section - 3.1.2 Semantic distance acquisition).

2.1.1.2 Upper ontologies

Similar to domain ontologies, upper ontologies tend to be manually generated. Unlike domain ontologies, an upper ontology (or foundation ontology) is a model of common or generic objects applicable across many domains. Consequently, they are principally used to either infuse domain ontologies with general concepts or to cater for interoperability between systems. There exist several standardised upper ontologies (e.g., The General Formal Ontology (GFO 2007), Suggested Upper Merged Ontology (SUMO 2006), CIDOC Conceptual Reference Model ontology (Doerr 2003)). Furthermore, at the time of writing, work was being undertaken to create a single standard universal upper ontology by the Standard Upper Ontology Working Group (SUO 2003).

One of the best known examples of an upper ontology is provided by WordNet (Miller, Beckwith et al. 1990). WordNet is a manually generated structure that defines the meaning of and the semantic relations between common objects (or rather their names). Consequently, although WordNet can only provide for a general level of understanding, it has been used in a number of related approaches (for example (Stoica and Heart 2006), (Resnik 1995), (Voorhees 1993)), as even a general level of understanding can provide for intelligent

search. For example, ArtEquAKT (Kim, Alani et al. 2002) utilises an ontology created by combining sections from Conceptual Reference Model (Doerr 2003) and WordNet (Miller, Beckwith et al. 1990) ontologies. This fused ontology allows ArtEquAKT to identify general knowledge fragments consisting of not just entities but also the relations between them. For example, the fact that 25/10/1881 is Pablo Picasso's date of birth can be recognised. Nonetheless, the domain of Painters studied as part of the ArtEquAKT project is much less volatile than that of computer science. Additionally, similarly to any approach relying on a manually generated ontology, ArtEquAKT is limited to concepts defined within that ontology.

2.1.1.3 Taxonomy for domain representation

Although WordNet (Miller, Beckwith et al. 1990) is one of the largest representations, it does not model everything. The amount of time required to manually categorise and model everything (assuming a team of experts that knows everything about everything can be found) would be astronomical. As a result, manually generated ontologies, and approaches that rely on them, are unlikely to ever be fully representative of their interest domain. Nevertheless, a specialised case of an ontology, a taxonomy, is easier to generate and, as a result, apply within the search domain. This is the case because taxonomies can be used to model only the relational structure, a hierarchy, of the domain or document collection. Consequently, due to the ease of implementation and hierarchical navigability providing an ability to browse the model at a desired level of granularity, taxonomies are commonly used with related search approaches (e.g., (Gligorov, Aleksovski et al. 2007), (Erofeev and Giacomo 2006), (Vogel, Bickel et al. 2005)). One of the best known examples of a taxonomy application within the search domain is provided by the Yahoo! (Grobelnik and Mladenic 1998) search engine. Nevertheless, Yahoo! taxonomy is of a manual nature and is, therefore, difficult to maintain and update.

However, an alternative, automated, approach of taxonomy generation can be provided by means of data clustering.

2.1.2 Data clustering

Data clustering is a technique of statistical data analysis which is typically regarded as a form of unsupervised ML *or unsupervised classification of patterns* (Jain, Murty et al. 1999). Data clustering is achieved through *division of data into groups of similar objects* (Berkhin 2006) – clusters. Consequently, effectiveness of a clustering approach depends on the type and quality of the criteria used to determine similarity of objects to be clustered. Although a review of different weighting approaches is beyond the scope of this thesis (see (Jain, Murty et al. 1999) for a detailed introduction), most are based on Euclidean distance calculations of vector/matrix document representations or variations thereof (e.g., tf.idf (Salton 1971) in VSM). Because Euclidean similarity values are derived from mathematical approximations, they do not necessarily represent a real measure of similarity between words. Consequently, this thesis proposes an improvement to an existing algorithm – NGD (1) (Cilibrasi and Vitanyi 2004; Cilibrasi and Vitanyi 2007) which is used as a means of the semantic measure of similarity detection for cluster generation.

From a search engine perspective, a model representation of the document collection can be created by means of clustering (e.g., (Koshman, Spink et al. 2006), (Ferragina and Gulli 2005), (Buntine, Lofstrom et al. 2004)). However, the type of model depends on the type of algorithm used: data clustering algorithms can be loosely classified as hierarchical or partitional.

2.1.2.1 Hierarchical and partitional data clustering

Hierarchical data clustering algorithms (Johnson 1967) find new clusters using existing ones. An algorithm starts by initiating all items as a separate cluster (agglomerative clustering) or a single cluster containing all elements (divisive clustering) and proceeds to iteratively merge (agglomerative) or split (divisive) clusters based on a measure of similarity or dissimilarity. This process continues until a specified number of clusters is reached (or no further allocation is possible, for example, due to a similarity threshold) generating as an end result a relation hierarchy or a tree of clusters.

Partitional data clustering algorithms (Salton 1971) also use a measure of similarity or dissimilarity, however, all clusters are determined at once. The algorithm starts by, for example, generating or selecting a specified number of clusters and then proceeds to assign elements based on the similarity measure to these clusters. Consequently, although elements within the generated clusters are related, a relation structure between the clusters is not defined.

2.1.2.2 Clustering for domain representation

Partitional algorithms are efficient when it comes to speed of execution. As a result, they can be applied to large document collections or where speed of execution is of importance. However, they do not provide domain relation models, generated clusters are not related, and, thus, cannot be used for navigation purposes. Furthermore, automatically creating meaningful headings for generated cluster groups is difficult. Although there are technological limitations, the primary reason for the difficulty is the fact that different users have a different approach to grouping information (Kural, Robertson et al. 2001). Consequently, it is not feasible that a single, i.e. programmer defined, approach would suit everyone. Nonetheless, partitional algorithms are successfully applied

to group search results (e.g., (Koshman, Spink et al. 2006), (Clusty 2004), (Buntine, Lofstrom et al. 2004)). One of the most well known examples of such application is Carrot2 (Stefanowski and Weiss 2003).

Carrot2 is able to interface with a general search engine (for example Google or Yahoo!) to perform the actual search and then clusters returned results automatically. However, Carrot2, as well as any other approach based on an automatic partitional clustering algorithm, is not able to create meaningful and complete headings for cluster groups. Making automatic decisions of what names to choose to represent clusters is difficult, although this limitation is partially addressed by approaches like Flamenco (Flamenco 2007).

Flamenco - *Flexible information Access using Metadata in Novel Combinations* - project, originated at Berkley University of California, is a search interface framework designed to generate meaningful metadata hierarchies sorted into categories using Castanet algorithm (Stoica and Heart 2006). Flamenco's hierarchy generation is based on WordNet (Miller, Beckwith et al. 1990) and nearly-automated metadata hierarchy creation (Stoica and Hearst 2004) whereby suggestions for cluster names are automatically generated and grouped but are then presented to experts for selection. Although this approach is better at generating meaningful headings its scope and applicability are limited due to its reliance on WordNet and expert input. Furthermore, although improvements to cluster heading representational quality of a partitional algorithm can be achieved, the fact that unrelated cluster groups are generated is a limitation of partitional algorithms that cannot be addressed.

From a search engine perspective (for example (Grobelnik and Mladenic 1998) Yahoo!), search results, or document collections, are best represented using a

navigable hierarchy. A hierarchy carries information about the domain and, as a result, can be used to explore or browse at a chosen level of granularity hence allowing the user to discover previously unknown information. Such a hierarchy can be automatically generated using a clustering algorithm (for example (Cutting, Karger et al. 1992) Scatter/Gather). However, due to the time consuming iterative nature, research on automated hierarchical document collection taxonomy generation, an approach proposed in this thesis, is relatively scarce (Qi and Davison 2009). In contrast, automatic assignment and clustering of search result is relatively well addressed (e.g., (Koshman, Spink et al. 2006), (Clusty 2004), (Buntine, Lofstrom et al. 2004), (Wibowo and Williams 2002), (Peng and Choi 2002), (Chen and Dumais 2000)).

2.1.3 Summary

This section has introduced a number of relevant related concepts and approaches corresponding to stage 1 of the intelligent search. Specifically, section 2.1.1 introduced ontologies and taxonomies and section 2.1.2 introduced data clustering. Furthermore two research questions were identified and placed within context:

- the applicability of hierarchical clustering algorithms to large document collections
- the clustering measure of semantic similarity

2.2 Stage 2: The representation of the user's information need

This section describes general concepts and approaches related to stage 2 of the intelligent search – representation of the user's information need to facilitate search. However, it should be noted that applicability of many approaches on their own or to large document collections such as the WWW is limited. Consequently this thesis proposes a novel approach of handling user queries (see section 3.3 - SemaCS search and personalisation): the query itself is treated in the same way as documents in the collection and the interpretation (although currently not required due to the topic specific implementation) can be provided through data-based personalisation. As a result, this section should be viewed as an introduction or a background of general related concepts: section 2.2.1 defines DM; section 2.2.2 describes ML; section 2.2.3 describes NLP; finally section 2.2.4 introduces the Semantic web.

2.2.1 Data Mining

DM, also referred to as Knowledge Discovery in Databases (KDD), is used to discover previously unknown patterns and knowledge and/or for prediction. DM techniques are generally applied to structured datasets and have been defined as the *nontrivial extraction of implicit, previously unknown, and potentially useful information from data* (Frawley, Piatetsky-Shapiro et al. 1992) as well as *the science of extracting useful information from large data sets or databases* (Hand, Mannila et al. 2001). Consequently, because DM techniques can be automatically applied to large volumes of data and provide a means to discover patterns and knowledge (e.g., classification, association rule mining, clustering) they can also be defined as unsupervised ML (see next section).

2.2.2 Machine Learning

ML can be thought of as a subset of AI, though it also has close ties to DM and statistics (Witten and Frank 2005). In basic terms, ML is a set of algorithms and techniques that allow computers to discover information about the domain of interest. Such learning is generally accomplished by using statistical methods over large datasets and tends to include some level of human input/interaction known as supervised learning. ML can also be fully independent of any sort of human interaction; this type of learning is known as unsupervised. Neural networks (Haykin 1994) are one of the best known, or perhaps the most fascinating, of such ML approaches.

Neural networks are trainable systems capable of learning in either supervised or unsupervised manner. Such networks consist of interconnected processing elements (representing brain neurons) that work together to produce an output. Output of a neural network relies on cooperation of individual elements within that network. Consequently, overall function can still be performed even if some neurons stop functioning. A further example of supervised ML is provided by General Architecture for Text Engineering - GATE (Cunningham, Maynard et al. 2002) which implements an SVM learning algorithm (Y. Li and Cunningham 2005).

Although many other ML approaches exist (and SemaCS can be considered one of them) most cannot be deployed without training. Consequently, it is unlikely that any currently available approach can be appropriate in all situations, *the universal learner is an idealistic fantasy* (Witten and Frank 2005).

2.2.3 Natural Language Processing

NLP studies problems of automated generation and understanding of natural text (Manning and Schütze 1999). NLP is closely related to other areas of research and is rarely used on its own but generally forms a part of a system or a larger application that makes decisions and reasons over information extracted using NLP algorithms. NLPs closest relation is to AI and Ontologies simply because NLP requires extensive knowledge about the outside world to infer meaning. Similarly, this is the case with *Semantic-Based Approach to Component Retrieval* (Sugumaran and Storey 2003) which integrates a domain ontology in combination with a NLP query translator. Although this is a good example of a NLP application, it is applied to user query processing only. Furthermore, the reliance on a manually generated domain representation means the effectiveness, scope and accuracy of the approach are limited. Another example of NLP use is provided by Powerset (Powerset 2008) search engine. However, applicability of NLP to user information need interpretation is generally constrained. This is the case because WWW search queries generally consist of only a few words (for example, in (Koshman, Spink et al. 2006) a majority of queries consisting of 2 terms is reported). As only a few words are provided (for example, 'game download') they are not representative of natural language. Additionally, applicability of NLP to large document collections, due to processing overhead, is similarly restricted.

2.2.4 The Semantic web

The Semantic web is thought of as the next generation of the WWW (sometimes referred to as Web 2.0) and it contains information that can be directly processed by machines. In simple terms, the semantic Web is a framework created to make web pages easily understandable (meaning is accessible to both machines and people); it is based on description tags (Meta data) defined through ontologies, Extensible Markup Language (XML) (W3C 2004b), and RDF

(W3C 2004a). Semantic Web is also regarded as a type of weak AI. The concept of machine-understandable documents does not imply AI which allows machines to comprehend human concepts. However, the Semantic web does indicate a machine's ability to solve a well-defined problem by performing a set of well-defined operations on a set of well-defined data.

Almost every web page contains Meta data and most search engines can already process it. Meta data can be analysed and searched automatically because its semantic meaning is understood (or, at least, well-defined). This is achieved through the use of a dictionary of domain terms containing domain concept definitions and their meaning – a domain ontology. Though, it is still likely to be a while until Tim Berners-Lee's vision can be implemented:

> *[I] have a dream for the Web in which computers become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize.* (Berners-Lee and Fischetti 1999).

Nevertheless, there have been significant advances bringing this vision closer, for example, Swoogle (Ding, Finin et al. 2005), RDF (W3C 2004a) based news feeds, interactivity and customisation defined through agents and standards as well as novel approaches to searching in general. One of such new paradigms is Wolfram|Alpha answer engine (Wolfram 2009), which no longer treats user queries as text but as computational units (questions). This thesis, through definition of SemaCS framework, mNGD (2), domain hierarchy representation,

etc. (see chapter 3 - SemaCS: Semantic Component Selection) also contributes to this domain and this vision.

2.2.5 Summary

This section has provided an introduction of general concepts related to stage 2 of the intelligent search: DM, ML, NLP and the Semantic web. This section has further defined the placement of this thesis within the general related field:

- As SemaCS defines a clustering approach of document collection representation, it can be considered a statistical unsupervised approach of DM and ML.

- Similarly to NLP, SemaCS deals with user queries and document collections represented using natural language.

- Furthermore, like the Semantic web, SemaCS defines a machine understandable representation of the domain of interest. SemaCS further applies this representation to provide for stage 3 of the intelligent search described in the next section.

2.3: Stage 3: The comparison of the two representations

This section describes approaches and concepts corresponding to stage 3 of the intelligent search - a comparison between representation of the document content and user's information need, as characterised by search engines. However, it should be noted that search engines generally address all 3 stages of the computer-based intelligent search. Consequently, although described in this section, search engine relevancy extends beyond stage 3. The remainder of this section is structured in the following way: section 2.3.1 introduces software component portals and describes a number of closely related concepts and approaches; section 2.3.2 defines IR intelligent search, finally section 2.3.3 introduces personalisation.

2.3.1 Search and software components

A search engine is an IR system designed with a single purpose of finding information within a database (be it on a corporate network, the WWW or a personal computer). However, this term mostly refers to systems and approaches that locate information on the WWW. Typically, a WWW search engine would accept a query and return a list of results ordered by relevancy, date, location, etc. (e.g., Google). However, the need to search is not limited to the WWW. Further to general WWW search engines there are many dedicated or topic specific approaches created to function within one specific domain. Although this could be viewed as a limitation of applicability, it does guarantee higher result accuracy within the intended domain of application (even if by excluding all other domains from consideration) as well as an ability to address domain specific requirements. This thesis was motivated by the need of one such domain – that of software components.

Software component portals are searchable repositories containing software component descriptions. Because all search systems share some basic functionality, approaches and concepts described in previous sections are just as applicable here. However, component portals also deal with a host of specialised issues such as versioning, privacy, security, etc. An example of such a specialisation is provided by *COTS products characterization* (Torchiano, Jaccheri et al. 2002) which defines a description framework consisting of domain specific characteristics (for example, for database components response time is important) as well as a set of generic quality characteristics (resource utilisation, stability, maintainability, etc.) obtained from the ISO/IEC 9126 standard (ISO/IEC 2001). However, COTS products characterization domain taxonomy as well as software component categorisation and assignment are manually performed. Although many software component portals have moved towards automated solutions, many still rely on basic keyword-based search algorithms. Another example of such a basic approach is eCots (Mielnik, Lang et al. 2003) which provides COTS software component identification, characterisation, experience feedback, discussions and rating services. eCots also uses a manually generated domain taxonomy and manual assignment of software components to that taxonomy. Manual assignment and taxonomy generation can provide high levels of representativeness and accuracy. Nevertheless, any approach that relies on expert input does not scale. As a result a number of approaches were proposed to address this limitation:

SemRank (Anyanwu, Maduko et al. 2005) attempts to detect the possibility of a certain type of relationship to be more likely (conventional) or less likely (unconventional) in order to decide what to search for. However, SemRank relies on limited corpora to make these decisions; as a result its usefulness is similarly limited. This limitation is partially addressed by MUDABlue, another interesting example of an automated solution. MUDABlue (Kawaguchi, Garg et al. 2006) is able to infer software component function by looking at variable

types instead of fixed corpora. Although novel (and possibly applicable within open source community) MUDABlue relies on access to the source code and is therefore not applicable in a situation where only textual descriptions are available (as is the case with COTS software components). Yet another approach able to address this COTS specific limitation is *Web-based component evaluation* (Barbier 2004) which defines a software component testing interface. Using this interface it is possible to obtain information like manufacturer, performance, security constraints, required interfaces etc. and, as a result, removing the need to access source code. However, this framework relies on vendors implementing a standard interface. At present, there are no components complying with this approach apart from those used in one of the case studies.

2.3.2 Intelligent search

Although, currently, best known WWW search engines are still 'basic' (e.g., Google), some of the newer generation search engines are gaining in popularity (e.g., (Wolfram 2009), (Powerset 2008), (Ding, Finin et al. 2005), (Ferragina and Gulli 2005)). These search engines require a greater understanding of semantics i.e. meaning of words. As words may have different meaning (for example, as introduced in section 2.1.1.1 - Domain ontologies, word 'card' could refer to a 'playing card' or a 'graphic card'), their intended meaning in a given situation has to be established, this process is known as word meaning disambiguation. However, word meaning disambiguation approaches are either (Yarowsky 1995):

- Unsupervised – these generally achieve lower result accuracy but do not require direct manual input, instead ontologies or dictionaries are used e.g., WordNet-based (Voorhees 1993). However, applicability of such approaches directly depends on type and size of the ontology or dictionary.

- Supervised – these generally achieve higher result accuracy but require manual input e.g., collaborative tagging-based (Yeung, Gibbins et al. 2008). However, applicability of such approaches is severely limited due to the cost and amount of input required.

As a result, the most overall applicable scalable solution can be achieved through the reuse of large existing sources of expert input (for example information available on the WWW) to create an unsupervised approach that can be applicable within any domain – SemaCS is such an approach (see Chapter 3 - SemaCS: Semantic Component Selection).

SemaCS does not require direct manual input, instead input is statistically acquired from the WWW, or other available resources, and the document collection being searched. SemaCS is able to statistically analyse available data in order to generate relation scores (semantic distances) used in decision making and meaning disambiguation processes.

2.3.3 Search personalisation

Personalisation can be thought of as the process of detection and provision of relevant, to given person, content and tends to be a part of any web portal/homepage (for example Yahoo!) as well as search engines (for example

Google). However, a large drawback of personalisation is the fact that it requires information about the user; it is the way such information is acquired that poses a problem. Personalisation is largely, as far as search engines are concerned, implemented based on analysis of transaction logs and past searches (Mobasher, Cooley et al. 2000). Although statistical analysis of historical information is useful in determining overall trends in user behaviour (DM), its usefulness is limited when it is considered on the scale of a single search instance. Because someone is interested in IR does not imply that search results should always be biased towards IR topics. This shortfall of personalisation and language model approach of IR (Ponte and Croft 1998) has motivated the development of a novel SemaCS personalisation approach called 'data-based' personalisation (see section 3.3 - SemaCS search and personalisation). With data-based personalisation any type of textual information can be used to provide personalisation (e.g. requirements, documents, project documentation, publications, etc.). This concept has further resulted in a SemaCS search algorithm where searching is performed as an intersection between three specialised models: search, domain and description taxonomies. Though it should be noted that the use of ontology for personalisation purposes was considered before (e.g., (Ferragina and Gulli 2005), (Daoud, Tamine-Lechani et al. 2009)). However, structuring user profiles using data-based personalisation and mNGD (2) for semantic intersection is a novel approach.

2.3.4 Summary

This section has introduced a number of relevant concepts and closely related approaches. It has further outlined what motivated the development of SemaCS search and personalisation – a need for an accurate automated solution of selection, search and categorisation of software components. Additionally, this section has further identified and placed within context following research questions:

- Classification/understanding of user information need
- Personalisation as a means to define user information need

2.4 Conclusion

In this chapter a number of related concepts and approaches spanning a number of domains were introduced. This thesis touches upon DM, ML and NLP domains while being related to search engines, Semantic web, and utilises a taxonomy generated using a form of clustering. Therefore, although all of the areas and domains have been investigated, none of the areas were reviewed in as much a detail as could be expected from a thesis that could be placed within only one specific area of research. This chapter has further outlined research context, questions and the problem domain:

In section 2.1 a number of concepts and approaches corresponding to stage 1 of the intelligent search were introduced: ontologies, taxonomies and data clustering. This section has further identified the need for an automated hierarchical clustering algorithm applicable to large document collections. Consequently, this section has also identified and placed in context the need for a clustering measure of similarity that can be used with such an approach.

In section 2.2 general concepts related to stage 2 of the intelligent search were introduced: DM, ML, NLP and the Semantic web. This section has further defined the placement of this thesis within the related field. As a clustering approach of document collection representation, is defined (see section 3.2.1 - SemaCS hierarchical partitional clustering algorithm) SemaCS can be considered a statistical unsupervised approach of DM and ML. Similarly to NLP, SemaCS deals with user queries and document collections represented using

natural language. Furthermore, like the Semantic web, SemaCS defines a machine understandable representation of the domain of interest. SemaCS further applies this representation to provide for the 3 stages of the intelligent search.

In section 2.3 a number of relevant related concepts and approaches corresponding to stage 3 of the intelligent search were introduced. Specifically, search engines and personalisation. This section has further identified and placed within context a need for SemaCS search and personalisation approaches (see section 3.3 - SemaCS search and personalisation). The section has further defined SemaCS relation to search engines as well as the fact that search approaches generally implement a combination of approaches corresponding to the 3 stages of the intelligent search.

Chapter 3 SemaCS: Semantic Component Selection

The previous chapter related the 3 stages of the intelligent search to the 3 corresponding areas of research. Chapter 2 has further defined the placement of this thesis within the related field as well as the research context, questions and problem domain. In addition, the need for an automated hierarchical clustering algorithm applicable to large document collections was identified. Consequently, the need for a clustering measure of similarity that can be used with such an approach has also been identified. Finally, chapter 2 has identified a need for a flexible automated approach to software component categorisation and selection. SemaCS is intended to provide such an automated approach by utilising domain knowledge obtained independently from textual software component descriptions and the WWW – hypothesis $\mathbf{H_1}$. Consequently, to address research questions and the 3 stages of computer-based intelligent search identified in the previous chapter, a list of inter-related requirements which SemaCS must satisfy was defined:

- Text processing and semantic relation acquisition:
  - Text pre-processing (stages 1 and 2 of the intelligent search)
  - Semantic distance acquisition (stages 1, 2 and 3 of the intelligent search)
- Domain taxonomy generation and textual description categorisation:
  - Domain taxonomy definition and population (stage 1 of the intelligent search)
  - Textual description taxonomy (Taxpet) index generation (stage 1 of the intelligent search)
- Search and Personalisation:
  - Semantic query processing (stage 2 of the intelligent search)
  - Semantic search (stage 3 of the intelligent search)
  - Data-based personalisation (stages 2 and 3 of the intelligent search)

The remainder of this chapter is structured as a direct representation of the above requirements: section 3.1 describes SemaCS data access modules, text pre-processing and semantic distance acquisition algorithms; section 3.2 describes SemaCS domain taxonomy generation and textual description categorisation algorithms; and finally, section 3.3 describes SemaCS search and personalisation algorithms.

3.1 SemaCS text processing and semantic relation acquisition

In this section SemaCS data access, pre-processing and semantic distance acquisition algorithms are introduced: section 3.1.1 describes common parts of speech (POS) removal which takes place before semantic relation extraction and domain taxonomy generation; section 3.1.2 introduces NGD (1) algorithm (Cilibrasi and Vitanyi 2004; Cilibrasi and Vitanyi 2007) and proposed modification – mNGD (2); section 3.1.3 distinguishes SemaCS mNGD (2) from the only other existing NGD (1) modification proposed in (Gligorov, Aleksovski et al. 2007).

3.1.1 Common POS removal

Before SemaCS taxonomy generation can take place (see section 3.2.1 – SemaCS hierarchical partitional clustering algorithm) textual descriptions are extracted from the document collection. Upon extraction, textual descriptions are generally pre-processed to remove unrepresentative words and common POS (Berkhin 2006). However, any word can be relevant; relevancy depends on what is being searched for. Therefore, removing words from consideration is impractical. Consequently, no words are removed during this phase. Although this decision impacts negatively on overall efficiency, it also guarantees that no

information loss occurs. Nonetheless, common POS that carry no semantic information (such as: ‘a’, ‘as’, ‘is’, ‘it’, ‘the’, etc.) are removed.

POS removal is dependent on a list of predefined rules or learnable patterns (see section 2.2.2 – Machine Learning) and, although training or rule generation is a complex task, the removal process itself is not. Thus, it was deemed more efficient to employ an existing solution. ANNIE (a Nearly-New Information Extraction System (Hepple 2000)), a readymade POS removal and tagging solution provided as part of the GATE (Cunningham, Maynard et al. 2002) was used – see Figure 1.

```
START
    SemaCS      : read textual descriptions (extracted from document collection)
                : send unprocessed descriptions
    GATE        : receive unprocessed descriptions
                : for all descriptions
                      extract text
                      remove common POS
                : return pre-processed descriptions as word collection
    SemaCS      : receive pre-processed description word collection
END
```

Figure 1: GATE data pre-processing

3.1.2 Semantic distance acquisition

Core SemaCS functionality is dependent on its ability to detect similarity between words (semantic distance or relatedness). From these semantic distances a domain taxonomy is generated and user queries are matched to textual descriptions. SemaCS implements mNGD (2), a modified version of the NGD (1) algorithm (Cilibrasi and Vitanyi 2004; Cilibrasi and Vitanyi 2007) to

detect such distances between words. However, semantic relatedness measures can be broadly divided into two categories: statistical (i.e. distributional similarity) and those based on lexical resources (Budanitsky and Hirst 2006). As SemaCS does not rely on user input, lexical resource, or knowledge base it has little in common with lexical-based approaches (such as e.g. (Voorhees 1993; Budanitsky and Hirst 2006; Stoica and Heart 2006; Yeung, Gibbins et al. 2008)). However, SemaCS does employ mNGD (2), a statistical means of semantic relatedness acquisition. Thus, SemaCS is related to the two other most notable statistical approaches VSM (Salton, Wong et al. 1975) and LSA (Landauer and Dumais 1997). Although it should be noted that there exist many other statistical approaches – see (Budanitsky and Hirst 2006) for an introductory overview, see also (Baeza-Yates and Ribeiro-Neto 1999) for an overview of VSM and other classical approaches.

However, unlike the VSM, SemaCS mappings are based on mNGD (2) acquired semantic similarities not syntax. Relating to LSA SemaCS implements a much simpler approach that only performs basic calculations (i.e. only mNGD (2) mappings are used), no concept identification occurs nor are there any similarities in implementation. Nevertheless, SemaCS (due to simplicity) has a potential to be scaled to large document collections such as the WWW.

3.1.2.1 Google similarity distance

Google similarity distance is a metric capable of measuring semantic relatedness between words based on Kolmogorov complexity theory (Li and Vitanyi 1997). This metric is only able to detect that a semantic relation exists; it is not capable of detecting the type or kind of a relation that was detected (see (Resnik 1999) for further definition of semantic relatedness measures). Thus, NGD (1) and, consequently, mNGD (2) are not capable of providing for semantic search as is. Nevertheless, NGD (1) and mNGD (2) can be used by a semantic

search approach (i.e. a search engine) to index document collections and to interpret user information need (see section 2.3.2 Intelligent search). Similarly to the Probabilistic Theory of Relevance Weighting (Robertson and Jones 1976) the metric is purely statistical and thus has the potential to provide for a scalable fully automated solution. However, unlike the Probabilistic Theory of Relevance Weighting the metric is used without any kind of estimation (the actual word co-occurrence is acquired via the WWW while probability of occurrence of words within documents is not required at all). Similarly to LSA - Latent Semantic Analysis (LSA 2003) or VSM (Salton, Wong et al. 1975), NGD (1) relies on the fact that semantically related words co-occur more often than those that are unrelated. However, unlike LSA or VSM, NGD (1) does not rely on a small localised document collection, nor does it create a matrix of word co-occurrence/relations for all the words within the document collection. Although NGD (1) is not able to extract the actual meaning of the words on its own, it can be adapted to provide such functionality (Gracia, Trillo et al. 2006).

$$NGD\ (x,y) = \frac{\max(\log f(x), \log f(y)) - \log\ f(x,y)}{\log N - \min(\log f(x), \log f(y)} \quad (1)$$

For two words ***x*** and ***y*** Normalised Google Distance (NGD) is obtained by selecting most significant of either logarithm of occurrence in document collection of word **x** – $\log f(x)$ or logarithm of occurrence in document collection of word **y** – $\log f(y)$ and subtracting the logarithm of co-occurrence in document collection of words **x** and **y** – $\log f(x,y)$. The resultant value is then divided by logarithm of total number of documents in collection: $\log N$ minus the least significant of either logarithm of occurrence in document collection of word **x** – $\log f(x)$ or logarithm of occurrence in document collection of word **y** – $\log f(y)$.

Nonetheless, it should be noted that NGD (1), in its original form, requires implicit knowledge of the total number of pages **N** within its document collection (number of pages referenced by Google – as Google is used to acquire word frequencies). With WWW based calculations differences in **N**, given **N** remains fixed, do not have any significant impact on results (Cilibrasi and Vitanyi 2004) because overall changes to the number of referenced web pages, although significant in quantity, are not significant in statistical terms (when compared to the overall 'size' of the WWW). However, NGD (1) cannot perform as effectively in an environment where a significant number of documents changes rapidly (this is the case with personalisation based searches and/or company Intranets) as these changes may be significant in statistical terms. Consequently, resulting semantic distances would be likewise significantly affected. Additionally, with domain specific terms (which may be extremely rare in the public domain) even fractional real differences are significant. To address this limitation a modified version of the NGD (1) algorithm (able to function without **N**) was devised specifically for use with SemaCS – mNGD (2).

3.1.2.2 modified Google similarity distance

Although NGD (1) algorithm can be employed to detect similarity between words it requires implicit knowledge of the total number of pages **N** referenced by Google. Consequently, NGD (1) cannot function effectively in an environment where a number of documents is small and can change rapidly. Furthermore, fractional relatedness score differences caused by a discrepancy between real and assumed **N** values renders the significance weight of the NGD (1) generated score variant over time (see section 5.1.3 – Study result analyses). To address these limitations a modification of the algorithm was devised and called mNGD (2).

$$mNGD = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log\left(\frac{(\boldsymbol{f(x)} + \boldsymbol{f(y)})}{\boldsymbol{f(x,y)}}\right) - \min(\log f(x), \log f(y))} \quad (2)$$

Proposed mNGD (2) modification addresses NGD (1) limitation by removing dependency on **N** – $\log N$ which is replaced with a logarithm of the sum of occurrence in document collection of words **x** and **y** – $f(x) + f(y)$ divided by frequency co-occurrence in document collection of words **x** and **y** – $f(x, y)$. The impact and suitability of this modification was evaluated via a controlled experiment using both modified and unmodified version of the formula (refer to section 4.3 – Pilot study for experiment description).

3.1.2.3 Gligorov et al. Google similarity distance

To the best of our knowledge the only other **N** type modification to NGD (1) algorithm is proposed in (Gligorov, Aleksovski et al. 2007) where authors describe an ontology matching algorithm based on fuzzy dissimilarity *'sloppiness value'*. Authors further introduce a modification to NGD (1) that, similarly to mNGD (2), removes NGD (1) dependency on **N**. Gligorov et al. logically define their modification of NGD (1) as related to the *special form* of their queries: whereby word collection **x** forms a subset of word collection **y** – this is then simplified to arrive at the final algorithm referred to as gNGD (3):

$$gNGD(x,y) = \log f(x) - \log f(y) = \log \frac{f(x)}{f(y)} \quad (3)$$

Assuming that **x** is a collection of words and that it always forms a subset of collection of words **y**, Gligorov et al. modification is logically sound. However,

this is not the case outside of their specialised domain as **x** cannot be a subset of **y** if both **x** and **y** are distinct words for which semantic relatedness score is to be acquired. Consequently, gNGD (3) cannot be applied in a SemaCS environment.

3.1.3 Data access

Due to being topic specific, another core SemaCS requirement is an ability to function within a limited environment such as a company Intranet. Although acquiring mNGD (2) distances from the WWW is likely to provide higher accuracy of detection, this is not always possible (for example, due to security restrictions) or needed (for example, when only project or company documentation are to be processed). However, regardless of the implied data source, SemaCS algorithms can be applied without any modifications due to modularity of prototype design. SemaCS data access module is readily interchangeable without having an impact on other modules – see Figure 2. A total of 3 interchangeable data access modules were implemented for evaluation purposes: Google access, Yahoo! access, and Intranet access.

Google search engine was accessed by means of Java-based SOAP interface. However, Google enforces a daily query limitation of 2000 queries per 24 hours that made it unfeasible for study implementation (although, for consistency reasons, experiment described in section 4.3 Pilot study was performed using Google access module). Consequently, specialised Yahoo! HTTP access interface access module was created. Unlike Google, Yahoo! does not enforce any query limitations that made it an ideal candidate for use with SemaCS case studies. Finally, a local Intranet data interface module was implemented via MySql and Java ConnectorJ access driver.

```
START
    SemaCS          : get pre-processed word collection (Figure1)
                    : send word collection
    Data module     : receive word collection
                    : for all words in collection:
                        set rate (count) of occurrence
                    :return formatted word collection
    SemaCS          :receive formatted word collection
END
```

Figure 2: Rate of occurrence assignment

3.1.4 Summary

This section has introduced the rationale and design of SemaCS pre-processing and mNGD (2) semantic distance acquisition algorithms. Additionally, in this section a number of important implementation decisions related to SemaCS prototype implementation and data access modules were introduced. Furthermore, this section has distinguished SemaCS mNGD (2) algorithm from the only other existing NGD (1) modification proposed in (Gligorov, Aleksovski et al. 2007).

3.2 SemaCS taxonomy generation and description categorisation

The type of domain or document collection representation best suited for a search approach is of hierarchical nature (see section 2.1.1.3 – Taxonomy for domain representation). Such a model representation of the document collection, a taxonomy, can be created manually (for example (Grobelnik and Mladenic 1998) Yahoo!) or automatically by means of clustering (for example (Koshman, Spink et al. 2006), (Ferragina and Gulli 2005), (Buntine, Lofstrom et al. 2004)). However, the type of clustering algorithm used dictates the efficiency and the type of model generated. Furthermore, effectiveness of a clustering approach depends on the type and quality of the criteria used to determine similarity of objects to be clustered. Previous section defined SemaCS approach of similarity detection mNGD (2). This section defines SemaCS model generation: section 3.2.1 defines SemaCS hierarchical partitional clustering algorithm; sections 3.2.2, 3.2.3, 3.2.4 and 3.2.5 describe algorithm implementation stages 1, 2, 3 and 4; and finally section 3.2.6 describes textual description index generation.

3.2.1 SemaCS hierarchical partitional clustering algorithm

SemaCS categorisation and search depends on the quality of the domain representation (taxonomy). Consequently, to best represent the data being modelled, SemaCS domain taxonomy is generated automatically using contents from pre-processed textual descriptions and mNGD (2) – hypothesis $\mathbf{H_{1c}}$. However, SemaCS taxonomy is not a linguistic resource comparable to WordNet or any other 'standard' ontology. It should be noted that due to mNGD (2) limitations (see section 3.1.2.1 Google similarity distance), the type of links created are not comparable to those found in a manually generated taxonomy as the relations used to define the taxonomy are unpredictable and unknown. Nonetheless, SemaCS taxonomy does provide a means of categorisation, browsing (via mNGD relations) and emergent semantics. Such a model also has

a capacity to be representative of all textual descriptions it embodies as it is based purely on these descriptions. However, the type of algorithm used, directly dictates the type of model generated (see section 2.1.2.2 – Clustering for domain representation). Hierarchical clustering algorithms generate a hierarchical representation that can be browsed at a given depth or granularity level of the model (referred to as a 'tier' of the model from this point forward). However, due to the iterative time consuming nature of hierarchical algorithms, applicability to large document collections like the WWW is limited. On the other hand, partitional algorithms are efficient and execute quickly because all clusters are detected simultaneously. However, partitional algorithms do not create navigable cluster hierarchies.

The algorithm proposed as part of SemaCS is a hybrid hierarchical partitional clustering algorithm (see Figure 3). SemaCS algorithm implements the most basic steps of partitional clustering (to select most dissimilar representative top tier clusters) but then allocates to and sorts the contents of these clusters into related hierarchies.

START
initiate all elements as clusters
select most popular dissimilar unrelated clusters
form top hierarchy level - **tier1**
allocate all remaining clusters to most related **tier1** clusters
for all tier 1 clusters
select most popular dissimilar unrelated clusters
form second level hierarchy - **tier2**
allocate all remaining clusters to most related **tier2** clusters
END

Figure 3: SemaCS hybrid hierarchical partitional clustering algorithm

Furthermore, SemaCS clustering algorithm implements features that make it applicable to large datasets:

- unlike batch clustering where all documents must be present it can be used and updated at runtime (this is similar to (Ester, Kriegel et al. 1998) and (Sahoo, Callan et al. 2006))
- an approach similar to language modelling (Ponte and Croft 1998) is used to generate a global representation of the document collection; consequently elements of the document collection could be added or removed dynamically (as the model is not connected to specific instances)
- a tier limit (depth) controlling model complexity is introduced, thus allowing the model to be general and efficient enough to be applicable to large document collections (this is an adaptation of partitional clustering where a number of clusters k can be predefined).

Many of these ideas and approaches are commonly used, however, combination proposed in this thesis, as well as application of mNGD (2), are unique. The resultant SemaCS hybrid hierarchical partitional clustering algorithm definition (see Figure 3), is divided into 5 distinct stages:

- Stage 1: Tier 1 element identification (Figure 4)
- Stage 2: Element allocation to Tier1 parents (Figure 5)
- Stage 3: Tier 2 element identification (Figure 6)
- Stage 4: Element allocation to Tier2 parents (Figure 7)
- Stage 5: Textual description index generation (Figure 10)

Although interrelated, each stage can also be performed separately, for example, to add additional textual descriptions or update a specific tier. The remainder of this section is structured as a direct representation of these 5 stages.

3.2.2 Stage1: Tier 1 element identification

The first stage of domain taxonomy generation is identification of representative Tier 1 elements (see Figure 4). Tier 1 elements must be representative of the most general (common) concepts (with more concrete concepts forming Tier 2 and Tier 3 elements). Finding elements representing general concepts is a simple task as these tend to have the highest rate of occurrence. However, extremely common concepts were not allowed to form Tier1 elements. For example, 'ICT' is a generic concept and, within the computer science domain of interest, everything would be related to this concept hence resulting in a single Tier 1 element. It should be noted that hierarchical models do generally stem

```
START
   SemaCS        : get formatted word collection (created in Figure 2)
                 : for all words in collection:
                       if current word is not too rare for Tier1 and
                       if current word is not too common for Tier1 and
                       if current word length is > 3
                          send existing Tier1 elements and current word
   mNGD                   receive Tier1 elements and word to check
                          for all Tier 1 elements:
                             get Tier1 element relation to word
                                if average or stronger relation found
                                return true else return false
   SemaCS                 receive relation status of the word
                          If no average or stronger relation found (a new concept)
                             add word as new Tier1 element
                             remove word from collection
END
```

Figure 4: SemaCS Tier 1 element identification

from a single root element and, for future implementations of SemaCS, model generation algorithms could be adapted to allow this. However, currently this is not the case because such an approach would also require an unlimited number of tiers as the next most general concept would likely form a single Tier 2 element etc. Since the number of tiers does not have an impact on SemaCS accuracy, but does introduce a level of unneeded complexity, a simpler solution is employed. To provide an acceptable level of performance the number of tiers is limited to 3, and an arbitrary maximum rate of occurrence (on per study basis) is used to ensure a more uniform distribution of concepts within the model.

3.2.3 Stage 2: Element allocation to corresponding Tier1 parents

With Tier 1 elements identified (stage 1) all remaining words in collection are allocated to their corresponding Tier 1 parents (see Figure 5). However, at this stage, such allocation is transitionary as both Tier 2 and Tier 3 elements are allocated to corresponding Tier 1 parents (no decision concerning which elements would form Tier 2 or Tier 3 is taken).

```
START
    SemaCS      : get Tier1 elements (allocated in Figure 4)
                : get remaining word collection (from Figure 4)
                : for all words in word collection:
                      send Tier 1 elements and current word
    mNGD              receive Tier1 elements and word to assign
                        for all Tier1 elements
                            get relation to word
                        return strongest relation
    SemaCS            receive strongest Tier 1 relation
                      add word as child of Tier 1 element
                      move allocated word to transitionary collection
                      remove word from word collection
END
```

Figure 5: SemaCS element allocation to Tier1 parents

3.2.4 Stage 3: Tier 2 element identification

With transitionary Tier 2 and Tier 3 elements identified (stage 2) allocation of Tier 2 elements can take place (see Figure 6). A transitionary sub-list of child elements (on per Tier 1 parent) is extracted and Tier 2 elements are identified. Like with Tier 1, very common concepts are not allowed to form Tier 2 elements by using a maximum rate of occurrence limit.

```
START
    SemaCS      : get Tier1 elements (allocated in Figure 4)
                : get transitionary word collection (allocated in Figure 5)
                : for all Tier1 elements:
                      get child words from transitionary collection
                      for all child words:
                          if current word is not too rare and
                          if current word is not too common and
                          if current word length is > 3
                              get existing Tier2 elements (for Tier 1 element)
                              send Tier2 elements and current word
    mNGD                      receive Tier2 elements and word to check
                              for all Tier 2 elements:
                                  get Tier2 element relation to word
                                  if strong relation found
                                      return true else return false
    SemaCS                    receive relation status of the word
                              if strong relation to existing Tier2 element not found
                                  add word as new Tier2 element (new concept)
                                  remove word from transitionary collection
END
```

Figure 6: SemaCS Tier 2 element identification

3.2.5 Stage 4: Element allocation to corresponding Tier2 parents

With Tier 1 and Tier 2 elements identified (stages 1, 2 and 3) allocation of Tier 3 elements can take place (see Figure 7). This is the final stage of SemaCS

domain taxonomy generation algorithm where any remaining transitionary elements are allocated to corresponding Tier 2 parents.

```
START
    SemaCS        : get Tier1 elements (allocated in Figure 4)
    SemaCS        : get transitionary word collection (from Figure 6)
                  : for all Tier 1 elements:
                        get Tier2 children (allocated in Figure 6)
                        get child words from transitionary collection
                        for all child words:
                            send Tier2 elements and current word
    mNGD                receive Tier2 elements and word to assign
                        for all Tier2 elements:
                            get Tier2 element relation to word
                        return strongest Tier 2 relation
    SemaCS              receive strongest Tier 2 relation
                        add word as Tier2 child (new Tier3 element)
                        remove word from transitionary collection
END
```

Figure 7: SemaCS Tier 3 element allocation

With stage 4 of the algorithm complete the 3 Tiers are populated. A sample of these tiers is shown in Figure 8. This sample is extracted from the taxonomy generated as part of the University of Westminster case study (see section 4.5 Case study 2: University of Westminster SRS module search). The Tiers are stored in a simple MySQL table. Tier 1 elements, being the most general starting point of the taxonomy, only contain the word and rate of occurrence. Tier 2 and Tier 3 elements also contain a reference id and the strength of relation to their parent.

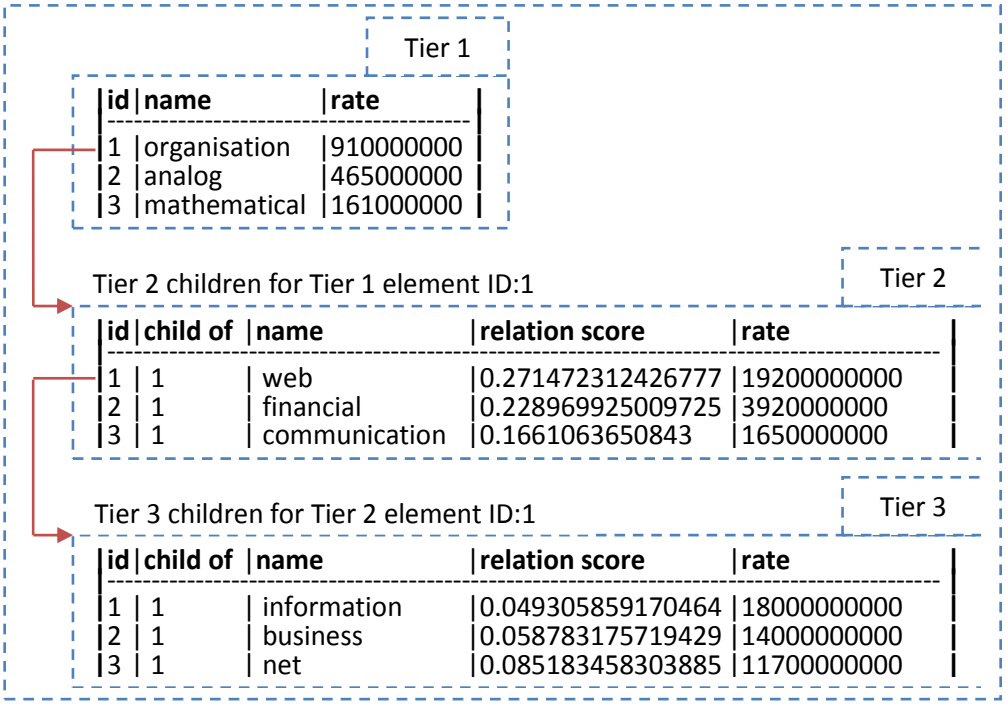

Figure 8: SemaCS taxonomy sample

3.2.6 Textual description index generation

To provide for flexibility (see next section) the SemaCS index generation stage was to form no direct 'hard' associations between the domain taxonomy and the descriptions. Instead taxonomy snippets (Taxpets) were created on per description basis (see Figure 9).

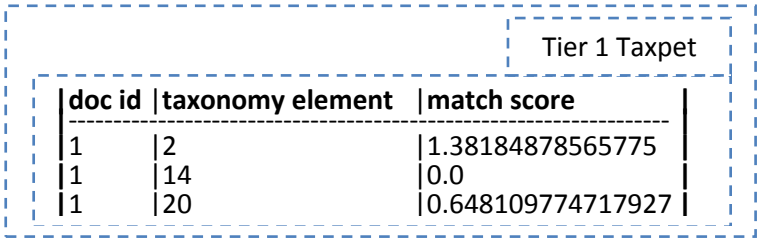

Figure 9: SourceForge.net study, Tier 1 Taxpet sample

These description Taxpets could then be dynamically intersected with domain taxonomy and search Taxpet (see section 3.3.1 – Semantic query processing and search) to perform a search based on semantic relation indexes between these taxonomies obtained via mNGD (2). A sample taxpet generated for one of the SourceForge.net case study software components (see section 4.4 Case Study 1: SourceForge.net) is shown in Figure 9. Taxpets can be created at any of the 3 levels of granularity (the 3 Tiers). This is particularly useful for an approach similar to query expansion as well as result weighting algorithms. However, Taxpet structure remains uniform regardless of granularity level.

```
START
    function getRelation (Tier elements and word to assign)
    mNGD            : receive Tier elements and word to assign
                          : for all Tier elements:
                              get Tier element relation to word
                          return strongest relation or 0

    SemaCS          : get Tier 1 elements from persistent store
                    : get description word collection (Figure 2)
                    : for all words in collection:
                          send getRelation(Tier1 elements and current word)
    SemaCS                receive strongest relation (or error)
                          if relation found
                              write relation score to Tier 1 element (TC1 relation)
                          else break
                          get Tier 2 child elements of identified Tier 1 element
                          send getRelation(Tier2 elements and current word)
    SemaCS                receive strongest relation (or error)
                          if relation found
                              write relation score to Tier 2 element (TC2 relation)
                          else break
                          get Tier 3 child elements of identified Tier 2 element
                          send getRelation(Tier3 elements and current word)
    SemaCS                receive strongest relation (or error)
                          if relation found
                              write relation score to Tier 3 element (TC3 relation)
END
```

Figure 10: Textual description index generation

Although simple, this method of maintaining association links is very flexible as both the model and descriptions can be updated independently of each other. However, due to resource limitations, a design decision has been taken to store all Taxpets in a single structure being an index of Taxpets to domain taxonomy relation (see Figure 10). It should be noted that this design decision did not have an impact on SemaCS accuracy or association mappings because both study data and domain representations were fixed (see chapter 5 – Result analyses).

3.2.7 Summary

This section has described SemaCS domain taxonomy and textual description index generation algorithms as well as important implementation decisions:

- SemaCS domain representation depth limit of 3 tiers
- Generation of Taxpets to domain taxonomy relation index

Although scalability to very large document collections (like the WWW) is beyond the scope of this thesis, SemaCS implementation could be further improved by modifying textual description index generation algorithms. Currently, indices are generated and stored as a single structure because this provides a sufficiently scalable solution for case study validation. However, 'real world' data is likely to evolve and change. As a result, a significant scalability improvement could be achieved by keeping such indices as separate entities (on per textual description basis as was originally planned) and interpreting their relatedness dynamically.

3.3 SemaCS search and personalisation

This section describes SemaCS search and Text-based search algorithms. SemaCS search could be described as an end result of an intersection of search taxonomy, domain taxonomy and description taxonomy. Each structure is a separate entity related to the others by means of semantic distances acquired via mNGD (2). Because each structure remains separate, each can also be updated dynamically without impacting others (for example, search taxonomy can be updated through data-based personalisation or RF, domain taxonomy can be updated by adding more descriptions).

This section further introduces SemaCS data-based personalisation approach. Personalisation is generally achieved (see section 2.3.3 – Search personalisation) through application of statistical information (such as past searches or preference settings). SemaCS provides search personalisation by processing any type of textual information (for example, publications, project documentation, extended descriptions) this was worded as hypothesis $\mathbf{H_{1d}}$.

The remainder of this section is structured in the following way: section 3.4.1 describes SemaCS query processing and semantic search algorithms, section 3.4.2 describes Text-based search algorithm, and section 3.4.3 describes SemaCS data-based personalisation.

3.3.1 Semantic query processing and search

All SemaCS queries are pre-processed to remove common POS; a search Taxpet is then generated from these pre-processed queries. Search Taxpet generation algorithms are identical to those used for description Taxpet generation (see section 3.2.6 – Textual description index generation). However,

once a search Taxpet representation of user query is formed, it could be expanded using relevant objects from user taxonomy or data-based personalisation. Although query expansion was part of the original design, it was not required to validate SemaCS within the chosen case study domains. Therefore query expansion has been left as further work.

With a formed search Taxpet SemaCS search is performed by intersecting, by means of semantic distances acquired via mNGD (2), description Taxpets and search Taxpet with domain taxonomy serving as an intermediary. Intersection and semantic distance – score – of that intersection are used to detect possible matches (see Figure 11). While this approach may not be processor efficient (when compared to Text-based search for example) it has the potential of being dynamically configurable and biased towards specific requirements without causing any system wide effects or impacting other users.

3.3.2 Text-based search

SemaCS Text-based search (see Figure 12) was created for evaluation purposes (a comparison search approach was required, see section 4.4 – Case Study 1: SourceForge.net). As it is very basic (and likely to be familiar) no thorough description is needed: a keyword such as ‘sat’ would match ‘Saturday’, ‘Satsuma’, ‘ASAT’ and so on. This algorithm searches all textual descriptions for an occurrence of each supplied keyword. If a match is found it is displayed (a greater number of matches within a single description results in higher score).

```
START
    SemaCS          : get Search Taxpet (Figure 7)
                    : send Search Taxpet to Data module
    Data module     : receive Search Taxpet
                    : for all search Taxpet entries:
                        get descriptions matching TC1 (match at Tier 1)
                        if match/es found
                            set description score to TC1 match score
                            add description to search result collection
                        else break
                        get descriptions matching TC2 (match at Tier 2)
                        if match/es found
                            set description score to TC2 match score
                            add description to search result collection
                        else break
                        get descriptions matching TC3 (match at Tier 3)
                        if match/es found
                            set description module score to TC3 match score
                            add description module to search result collection
                    : return search result collection
    SemaCS          : receive search result collection
                    : extract distinct descriptions from search collection
                    : for all distinct descriptions:
                        extract identical descriptions from search collection
                        for every occurrence in search result collection:
                            add match score to description total score
                    : sort search results - order of score
END
```

Figure 11: SemaCS search

3.3.3 Data-based personalisation

Searching and personalisation are closely linked as one affects the other. Information obtained from personalisation feeds back into the search process as RF. However, user-sourced information may not be available or difficult to obtain (see section 2.3.3 – Search personalisation). With SemaCS data-based approach, personalisation could be provided by means of any type of textual information (such as publications, project documentation, extended descriptions,

etc.) when required. This is achieved by extracting a textual description from, for example, project documentation provided by the user. This description is then processed (see section 3.2.6 – Textual description index generation) and used to infuse search Taxpet (see previous section) with additional information. Thus, personalising, search results towards the type of topic(s) discovered in the provided textual description.

```
START
    SemaCS          : send search query to Data module
    Data module     : receive search query
                    : for all keywords in query:
                          for all descriptions:
                              if keyword matches description
                                    add matched description to search results collection
                    : return search results collection
    SemaCS          : receive search results collection
                    : extract distinct descriptions from search result collection
                    : for all distinct descriptions:
                          extract identical descriptions from search result collection
                          for every occurrence in search results collection:
                              add 1 point to description score
                    : sort search results - order of score
END
```

Figure 12: SemaCS Text-based search

SemaCS data-based personalisation approach was originally motivated by the need to provide user level profile based personalisation. Thus it was designed to provide high levels of personalisation and, consequently, result relevancy (hypothesis $\mathbf{H_{1d}}$) without having any adverse effects on global domain taxonomy or other users. However, although profile based data personalisation forms an integral part of the SemaCS concept, only basic domain-wide data-based personalisation to evaluate hypothesis $\mathbf{H_{1d}}$ was implemented (see to section 4.5 - Case study 2: University of Westminster SRS module search). Nonetheless,

the algorithms implemented to provide domain-wide data-based personalisation are identical to those used to provide user level data-based personalisation. Thus, although due to resource limitations these algorithms could not be tested with actual users or user profiles, initial evidence supporting hypothesis $\mathbf{H_{1d}}$.was provided (see section 5.1 Pilot study).

3.3.4 Summary

This section described SemaCS search, Text-based search and domain taxonomy generation algorithms. It has further introduced SemaCS data-based personalisation concepts and defined important implementation decisions. Currently SemaCS does not implement user-level personalisation features because an evaluation of $\mathbf{H_{1d}}$ hypothesis could be achieved using domain-wide data-based personalisation only. As a result user-level personalisation has been left as further work.

3.4 Conclusion

This chapter elaborated on SemaCS design decisions and algorithms. This chapter has further described SemaCS prototype implementation addressing the 3 stages of computer-based intelligent search. Like the 3 stages of the intelligent search, SemaCS was divided into 3 distinct data-driven stages as well as 3 modules (see Figure 13).

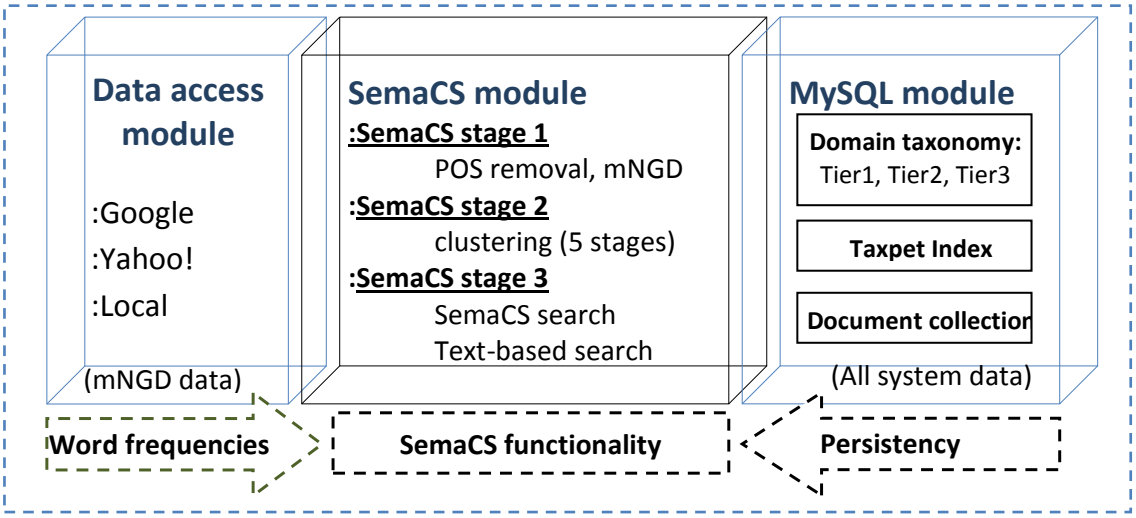

Figure 13: SemaCS modules and stages

Section 3.1 defined stage 1 of the SemaCS process:

Before being applied to Stage 2, textual descriptions are extracted from the document collection and pre-processed. However, SemaCS pre-processing only removes common POS, no words are removed. Additionally, SemaCS does not implement POS removal, instead an existing solution – ANNIE (Hepple 2000) is used. This section has further defined SemaCS semantic relation acquisition mNGD (2) algorithm, an **N** independent modification of NGD (1) algorithm. Finally, mNGD (2) application to word relation extraction by means of 3

interchangeable data access modules (Google access, Yahoo! access, and Intranet access) was defined.

Section 3.2 defined stage 2 of the SemaCS process:

SemaCS implements an automated hybrid hierarchical partitional clustering algorithm to generate a domain taxonomy. This algorithm is divided into 5 interrelated stages, which can also be performed separately to add additional descriptions or update a specific domain taxonomy tier. Additionally, this section defined SemaCS taxonomy generation algorithm implementation decisions: SemaCS domain taxonomy representational depth is limited to 3 tiers and SemaCS textual description Taxpets are stored in a single index.

Finally, section 3.3 defined stage 3 of the SemaCS process:

SemaCS search is a result of intersecting, by means of semantic distances acquired via mNGD (2), of textual description Taxpets and search Taxpet with domain taxonomy serving as an intermediary. This section has further defined a simple Text-based search algorithm created for evaluation purposes. Additionally, SemaCS data-based personalisation approach, provided by means of application of textual information to infuse a search taxpet, was defined. However, only basic domain-wide data-based personalisation to evaluate hypothesis $\mathbf{H_{1d}}$ was implemented.

Chapter 4 Methodology

The previous chapter has defined the 3 stages of the SemaCS algorithm: text pre-processing and semantic distance acquisition, domain taxonomy generation and textual description index generation, search and personalisation. It has further defined the SemaCS prototype created to provide a means of evaluating primary and secondary hypotheses and elaborated on design and implementation decisions.

This chapter introduces the chosen evaluation methodology. It defines the case study experimental approaches, and elaborates on study contexts, participants, data collection procedures and designs. This chapter further introduces statistical IR evaluation criteria of Precision, Recall, F-score, interpolated average Precision Recall and Mean Reciprocal Rank (MRR) chosen as a means to evaluate SemaCS.

The remainder of this chapter is structured in the following way: section 4.1 introduces the chosen evaluation methodology; section 4.2 defines chosen evaluation criteria; section 4.3 describes SemaCS feasibility study; section 4.4 describes SourceForge.net case study; and finally, section 4.5 describes University of Westminster SRS module search case study.

4.1 Methodology rationale

Case study methodology is a common means of evaluating IR approaches (Benbasat, Goldstein et al. 1987). Although, general evaluation could be achieved through a set of controlled experiments this is seldom the case where realistic in-use evaluation is sought. SemaCS is designed to provide for a user's

view of the world and to assess its effectiveness it should be evaluated against that view. This approach is referred to as operational and, although compared to controlled experiments it sacrifices a measure of observability and repeatability, it does provide a more realistic means of evaluation (Robertson and Hancock-Beaulieu 1992). This is our rationale for choosing an operational case study methodology. Unlike a set of experiments designed to test a few specific aspects over countless iterations, an operational case study makes it possible to analyse the end result in a realistic environment (although, generally, over a smaller number of iterations). These types of case studies require a large amount of time and effort therefore the design and implementation was careful and meticulous. A feasibility (pilot) study was first performed to suggest success. This pilot study was used to validate the methodology, the design and the SemaCS implementation as well as to provide initial evidence supporting primary hypothesis $\mathbf{H_1}$.

With the pilot study completed, two further case studies were designed. In part, this choice was made with the view to better evaluate the secondary hypotheses $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$, $\mathbf{H_{1c}}$, and $\mathbf{H_{1d}}$. However, the primary objective for performing two studies within two domains (software component descriptions and computer science module descriptions) was to provide the evidence with the view to better evaluate the $\mathbf{H_1}$ hypothesis. If disproven, $\mathbf{H_1}$ would render any secondary hypothesis meaningless. As both studies are distinct and based within different application domains, both are better suited to evaluate $\mathbf{H_1}$ hypothesis than either on their own.

4.2 SemaCS evaluation criteria

A means to assess the effectiveness and impact of the proposed approach is required. Assessment criteria are closely related to application domain and can

therefore differ greatly. As a result, choosing what to evaluate, as well as how, is of great importance. Furthermore, a question of result significance can be interpreted differently depending on the intended area of application. For example, within approaches designed for the WWW (such as web search engines) an ability to find all relevant documents (possibly regardless of the number of those that are irrelevant) is acceptable, while for intelligent applications (such as expert systems) the ability to find only relevant documents (with as few irrelevant as possible) is sought.

To evaluate SemaCS, a set of criteria to assess accuracy and impact on results is required. The SemaCS approach means assessment can only be based on queries supplied by study participants and query results returned back for participant appraisal. This indicates SemaCS has the ability to process and apply semantics extracted from the natural text being evaluated. Furthermore, chosen evaluation criteria had to provide a means to compare our study results to related work.

In view of the above requirements, an evaluation of SemaCS effectiveness was achieved by means of analysis of result accuracy and relevancy as well as the impact of personalisation on these results. Evaluation criteria were based on a combination of expert judgement (to compare against) and standard IR statistical measures of Precision (4), Recall (5) and F-Score (6). These criteria provided a means of direct statistical comparison of both studies, as well as a means to directly compare study results to related work, (as summarised by (Singhal 2001)) Precision and Recall are widely used and accepted within the IR community).

$$P = \frac{Rel}{Tot} \quad (4) \qquad R = \frac{Rel}{Av} \quad (5) \qquad F = \frac{2PR}{P+R} \quad (6)$$

Although Precision, Recall and F-Score are effective means of overall evaluation, they do not take into consideration the performance of an approach generating a ranked list of results. Thus two further evaluation techniques were used to evaluate SemaCS result ranking and relevancy criteria: 11-point interpolated average Precision Recall (Salton and McGill 1983) and MRR (Voorhees 1999).

4.2.1 Precision

Precision **P** represents the proportion of relevant terms or documents **Rel** to that of total number of returned terms or documents **Tot**. This represents a measure of performance evaluation with an ideal result being: all returned documents are relevant to the query.

However, with search approaches in general, it is a common modification to assume a ‘cut-off’ point for query results. A cut-off point is used to either provide the best possible relation between recall and precision (point of intersection) or to limit the number of returned documents (as some queries may result in an almost infinite number of matches). A search algorithm would then not return documents that are below a given cut-off score. This approach was employed with SourceForge.net study (see section 5.2.3 - Study result analyses).

4.2.2 Recall

Recall **R** represents the proportion of relevant terms or documents **Rel** to that of all available relevant terms or documents **Av**, with an ideal result being: all terms or documents relevant to the query are returned.

For SourceForge.net study a set of relevant documents was decided upon beforehand (see section 4.4.4 Study data). This was achieved by means of expert judgement and a fixed number of predetermined scenarios. However, second study presented more of a challenge as participants were allowed to search for any free choice module (as they would naturally, using University of Westminster Student Record System). This resulted in natural queries that are 'real' (instead of fixed by an expert). However, each such query had to be treated as a distinct case with all possible relevant matches decided on per case basis by a panel of experts.

4.2.3 F-Score

F-Score represents a weighted average of precision **P** and recall **R** with an ideal result being: all available relevant terms or documents (and only those that are relevant) are returned. Although Precision and Recall are important evaluation criteria, they do not necessarily provide a means of complete assessment of method being evaluated. This is the case because both measures can be 'biased'. It is possible to tune an algorithm to return only most relevant results (for example by using a high cut-off point) hence boosting Precision. Likewise, it is possible to tune an algorithm to provide a high measure of Recall for example (as an extreme case) by displaying the whole database. However, it is not possible to bias an algorithm to simultaneously provide high Precision and high Recall unless it is capable of doing so. As a result, a combined measure based on both Precision and Recall – an F-Score can be used to provide a single evaluation measure.

4.2.4 Interpolated average Precision Recall

For any system that returns an ordered list of results, Precision and Recall can be plotted on a graph to provide a Precision Recall curve. Such curves are a

means of evaluating and visualising the effectiveness of an approach at retrieving and ranking relevant documents and, consequently, of its perceived usefulness to the user. A result set with all the relevant documents at the end of a 100 page list is not as useful as a result set containing all the relevant documents on the first page. However, a Recall Precision curve is generated on per result basis and has a saw-tooth shape. To smooth the curve interpolated precision is used, to make the technique applicable as a means of overall evaluation an 11-point interpolated average Precision Recall curve (Salton and McGill 1983) is traditionally used (Manning, Raghavan et al. 2008).

Interpolated precision at any given recall point is defined as either the precision at that given point or at any further point along the recall scale where precision is higher. These interpolated precision values are measured/estimated at the 11 points of a recall scale (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). Finally, precision values from all the result sets are averaged to produce a single curve representation of the overall performance of an approach being evaluated.

4.2.5 Mean Reciprocal Rank

MRR (7) is a measure of evaluation used in the TREC question answering track (Voorhees 1999). Like the 11-point interpolated average Precision Recall curve, MRR evaluates the effectiveness of a result ranking approach. However, unlike an 11-point interpolated average Precision Recall curve, MRR provides a singular measure.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank\ i} \quad (7)$$

For any given query, the Reciprocal Rank (RR) is 1, divided by the position (rank) of the first relevant element in the result set. If there is no relevant document within the top five elements of the result set, the RR for that result set is taken as zero. RR values from all the result sets are averaged to produce the MRR, a representation of the overall effectiveness of an approach to rank the relevant results.

4.2.6 Summary

Precision, Recall and F-Score criteria provide a means of overall evaluation. However, these evaluation measures are not designed to effectively assess an IR approach capable of generating ranked answers. As SemaCS is such an approach two further evaluation techniques are used to assess its effectiveness: MRR and 11-point interpolated average Precision Recall curves.

Though, additional measure of importance can be placed on other attributes such as execution performance or domain representation. Execution performance is, however, beyond the scope of this thesis mainly because it is largely dependent on external factors (such as the deployment platform, network connection, implementation and load). Furthermore, issues of execution performance are relative to application domain and purpose. Likewise, it was decided not to evaluate SemaCS domain taxonomy and index structures individually because of their inherent nature; these models are generated via automated means and are used to perform automated reasoning. Consequently, their combined ability to represent the domain of interest, infer user information need and, as a result, return relevant textual descriptions is evaluated.

4.3 Pilot study

This section describes a feasibility study implemented to provide initial evidence regarding $\mathbf{H_1}$ hypothesis. This study was further intended to aid with focus and design of SourceForge.net (section 4.4) and University of Westminster Student Record System module search (section 4.5) case studies.

4.3.1 Pilot study experiment: Semantic distance detection NGD vs. mNGD

The ability to detect a degree of relatedness between words is the key feature of the SemaCS approach. As a result, semantic distance detection is the first step in validation.

To evaluate suitability and performance of mNGD (2) and NGD (1) generated relation scores, a set of 31 word pairs (see Table 2) with associated relatedness scores (statistically generated from input provided by a group of 51 students) were acquired from the Rubenstein and Goodenough 1965 study. The data formed a gold standard against which performance of both algorithms was evaluated.

A total of two experiments were performed (see section 5.1 - Pilot study), both on identical data and within an identical environment. The first experiment was performed using mNGD (2) algorithm, while the second experiment was performed using original NGD (1) algorithm. These two experiments were designed with a view to provide evidence to either prove or disprove hypothesis $\mathbf{H_{1b}}$. (see section 5.1.3 - Study result analyses).

#	Word1	Word2	Score	#	Word1	Word2	Score
1	cord	smile	0.02	**16**	cord	string	3.41
2	rooster	voyage	0.04	**17**	glass	tumbler	3.45
3	noon	string	0.04	**18**	grin	smile	3.46
4	fruit	furnace	0.05	**19**	serf	slave	3.46
5	autograph	shore	0.06	**20**	journey	voyage	3.58
6	automobile	wizard	0.11	**21**	autograph	signature	3.59
7	mound	stove	0.14	**22**	coast	shore	3.60
8	grin	implement	0.18	**23**	forest	woodland	3.65
9	asylum	fruit	0.19	**24**	implement	tool	3.66
10	graveyard	madhouse	0.42	**25**	cock	rooster	3.68
11	glass	magician	0.44	**26**	boy	lad	3.82
12	boy	rooster	0.44	**27**	cushion	pillow	3.84
13	cushion	jewel	0.45	**28**	cemetery	graveyard	3.88
14	monk	slave	0.57	**29**	automobile	car	3.92
15	asylum	cemetery	0.79	**30**	midday	noon	3.94
				31	gem	jewel	3.94

Table 2: Pilot study Experiment golden standard

4.3.2 Summary

This section described the SemaCS feasibility study. This study provided positive evidence supporting hypothesis $\mathbf{H_{1b}}$ as well as underlying principle approaches (see section 5.1 Pilot study). Therefore, this study has provided positive initial evidence supporting $\mathbf{H_1}$. It has further aided with focusing and designing the SourceForge.net and University of Westminster Student Record System module search case studies (described in next section). Furthermore, a number of problems with the SemaCS algorithm implementation were identified and corrected, hence avoiding any negative impact on the following case studies.

4.4 Case Study 1: SourceForge.net

SourceForge.net (SourceForge.net 1999) was chosen as a subject for this study because it provides a means to evaluate the primary and secondary hypothesis (see section 5.2.3 - Study result analyses) within the context of a large and well known software component portal. Additionally, SourceForge.net has been used in a number of related work evaluations (e.g. (Kawaguchi, Garg et al. 2006) and (Vanderlei, Durão et al. 2007)).

4.4.1 Research principles

A mix of subjects with experience ranging from very limited (Biosciences) to moderate (Computer Science undergraduates) to advanced (researchers) were asked to participate. In every case participation has occurred on a voluntary basis with a total of 40 participants taking part in this study. There are no ethical implications as the only statistical data required for Precision, Recall and F-Score calculations were recorded; no record of any personal information (such as age, name, gender, etc.) was made. None of the participants had been given prior knowledge concerning their task or this study. A standard uniform description of a task and of the purpose of the study was provided prior to participation taking place (see Appendix A4). Participants had an option to withdraw from the study at any point during their participation. A further option of obtaining study results (once compiled) was made available.

4.4.2 Study participants

To allow for variation and to ensure minimal bias of results a total of 40 subjects from a variety of occupational backgrounds participated in the study. These participants were randomly selected for voluntary participation with no preference for age, gender or experience. Due to random variation of

backgrounds and previous experience (and the fact that participants were not aware of tasks used) it is possible to generalise that any greater number of participants would not have any significant impact on results (assuming experimental setup and test cases are to remain unmodified and participants unbiased).

Participants were from one of two general groups: undergraduates and researchers. Researchers participating in the study were Electronics students. Undergraduate students participating in this study were from biosciences, computer science and electronics courses. All students had participated on a voluntary basis. Additionally, both groups of participants had a low degree of familiarity with the assigned tasks (though, due to professional background, some familiarity could be assumed). Although none of the participants were in any way aware of this research or of its test platform, some general familiarity with the process of searching (for example using Google) could be assumed.

4.4.3 Study environment and platform

Two platforms with the same interface (see Figure 14) were deployed within the same hardware and software environment: A Text-based search system and a SemaCS prototype implementation (see Chapter 3 - SemaCS: Semantic Component Selection). Both platforms were executed on a stand-alone Apache Tomcat 6.0 web server running on Samsung NP-Q1 UMPC. Prior to the case study a set of 51 software components was extracted from SorceForge.net corpora and a study domain taxonomy was generated. Once the case study started, the study domain taxonomy, environment and platforms have remained unmodified for the entire duration of the case study.

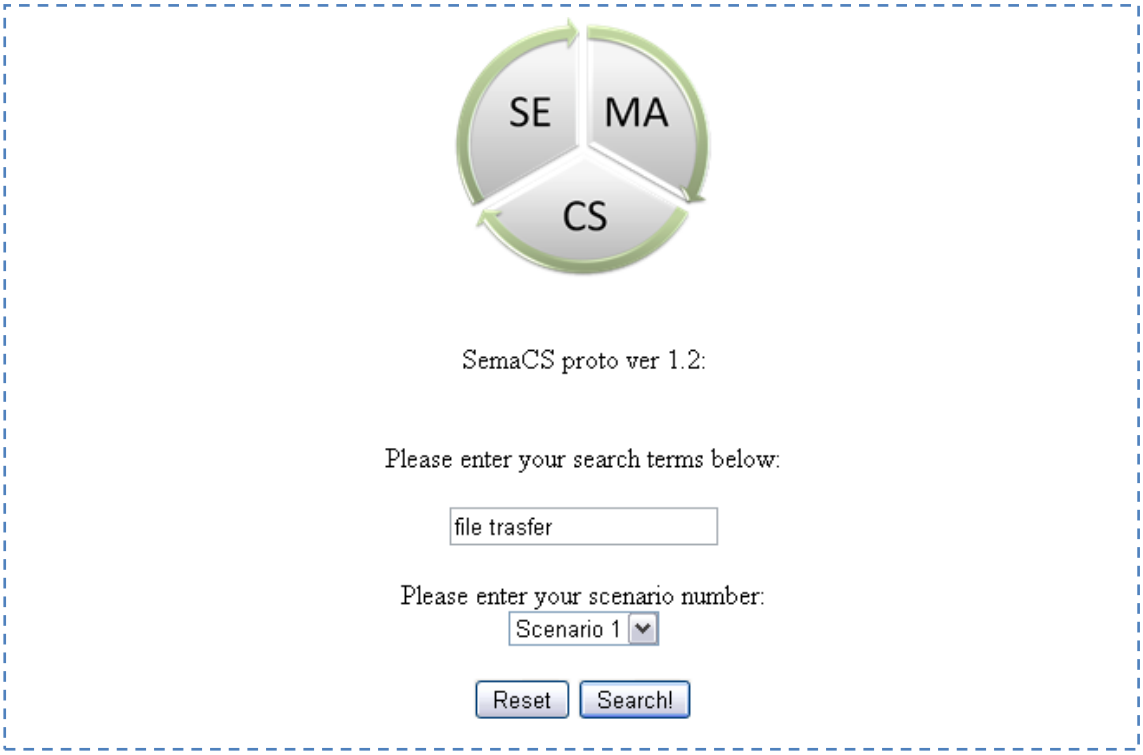

Figure 14: Case study 1 SemaCS search interface

4.4.4 Study data

For this study SourceForge.net software component database was used as a source to extract mNGD (2) semantic relation distances. SourceForge.net data (dated February 2008) were accessed by means of FlossMole (Howison, Conklin et al. 2004), an open-source project specialising in extraction of 'data dumps' from major software component portals including SourceForge.net (SourceForge.net 1999) and Freshmeat.net (Freshmeat.net 2002). A total of 130776 software component descriptions were extracted from the data dump and imported into a local MySql database (about 2% of the data was erroneous and could not be imported via automated means; these exceptions have been ignored).

Upon completion of the above import procedure, 51 software component descriptions for use in this study were randomly selected (see Appendix A1) from MySql copy of SourceForge.net data. The selection took place at a sample frequency of every 1000th software component with the starting point being varied randomly after every 10th component. Although a dataset of only 51 software components is a limitation, given prototype limitations and the demand on participants (participation was voluntary, consequently, subjects were not prepared to devote more than a few minutes of their time) a larger sample could not be used. Nonetheless, a dataset of 51 software components was deemed sufficient as proof of concept. Furthermore, a second study was designed (see next section) to collect further evidence. The remaining 130725 software component descriptions were used by SemaCS to acquire mNGD (2) semantic relation distances from (a replacement for the WWW). Five software components, chosen by the researcher, were used to create the five predefined scenarios; the remaining 46 software components provided the necessary noise data. Some of the noise samples turned out to be related to study scenarios (relevancy was decided on by the researcher before commencing the study). However, these unforeseen matches were not discarded because they guaranteed a more realistic measure of Precision and Recall (or at least a more realistic evaluation environment as real data is unlikely to be rigidly defined). A domain taxonomy was then generated based on these 51 components. Once generated, the domain taxonomy has remained fixed. As a result, the only variable in this study were the participant queries used.

4.4.5 Study data processing

mNGD (2) was used (see section 3.1.2 - Semantic distance acquisition) to process the data in order to generate indices and domain taxonomy. However, before indexing and model generation took place, case study data was pre-

processed using GATE (Cunningham, Maynard et al. 2002) to remove common POS and stop words.

4.4.6 Study data collection procedures

Only statistical data (see Table 3) required for Recall, Precision and F-score calculations were gathered. No record of any personal information was made.

Collected data	Description
Scenario Number	1 to 5 (see Table 5.3)
Participant's query	Keywords used as a search string
Time search started	System time search has commenced
Search results	Component Ids and scores
Result return time	System time results were displayed
Participant section	Component Ids (selected from returned results)

Table 3: Case study 1 automatically logged data (definition)

Data was collected automatically via an embedded logging facility with participant query and scenario number recorded via the search interface (see Figure 14) and participant selection(s) recorded via the result interface (see Figure 15). Although an option to select a match(es) was made available (via a check box within the results interface), these selection were not made use of for evaluation purposes. This is the case because scenarios, and consequently possible matches, were predetermined before the study commenced.

Figure 15: Case study 1 SemaCS results interface

4.4.7 Procedures

All study participants were assigned a single scenario to complete in order from 1 to 5. Once all 5 scenarios were performed the count reverted to scenario 1. Upon completion of the data collection all search queries were repeated using the opposite platform (for example, if the original query was performed using the SemaCS test platform then this same query was repeated using Text-based platform and vice versa) in order to make a direct statistical comparison between the two by excluding the human variable. That is, the only variation is the system being used with data, query, and human factors being identical and therefore negated. Five scenarios (correlating to the five distinct software components manually extracted from SourceForge.net) were designed for this study (see Table 4).

Scenario	Scenario description
1	Please find an application(s) that can be used to download and/or upload files to or from a server.
2	Please find an application(s) that can be used to monitor network traffic.
3	Please find an application(s) that can be used to remotely control a computer.
4	Please find an application(s) that can be used to play video and/or audio.
5	Please search for an application(s) that can be used to modify and/or convert sound files.

Table 4: SourceForge.net case study scenarios

A uniform explanation of this study as well as a description of the task to be performed was provided to every participant (Appendix A4) with participant query and results being automatically logged (see Appendix A2). A sample of the log is shown in Figure 16. To perform Precision, Recall, F-score and MRR

SemaSearch:: Scenario NO: 01; Request string: ftp upload software; Time search started: Thu Jul 10 12:25:55 BST 2008; Search term ftp matches T2: ftp; Score: 0.0; Search term upload matches T1: image; Score: 0.9439446237013213; Search term software matches T3: software; Score: 0.0; Matched components: ID: 24 Score: 12| ID: 8 Score: 7| ID: 29 Score: 7| ID: 6 Score: 6| ID: 27 Score: 6| ID: 43 Score: 6| ID: 47 Score: 4| ID: 16 Score: 3| ID: 49 Score: 3| ID: 42 Score: 3| ID: 48 Score: 2| ID: 31 Score: 2| ID: 35 Score: 2| ID: 18 Score: 2| ID: 46 Score: 1| ID: 2 Score: 1| ID: 38 Score: 1| ID: 11 Score: 1| ID: 4 Score: 1| ID: 36 Score: 1| ID: 1 Score: 1| ID: 45 Score: 1| ID: 13 Score: 1| ID: 41 Score: 1| ; Time results returned: Thu Jul 10 12:27:14 BST 2008; User returned results at: Thu Jul 10 12:29:17 BST 2008; User section: 24, 6,

Figure 16: SourceForge.net case study log sample

analyses, experiment log entries were imported into a Microsoft Excel spreadsheet on per query basis (a single line of the log representing a single

query). These queries were further split (using semicolons as a separator) into columns representing scenario number, user request (search) string, and a list of matching components returned by SemaCS. Imported data were then analysed using Precision, Recall, F-score and MRR formulas (see section 4.2 SemaCS evaluation criteria) to generate per query and total experiment average values shown in Appendix A3. These data were also used to generate per scenario and total experiment 11-point average Precision Recall curves (see section 5.2.3 Study result analyses).

4.4.8 Summary

SourceForge.net study was performed on two distinct systems: SemaCS and Text-based. All procedures were designed in such a way as to allow for removal of human variable in order to facilitate statistical evaluation and comparison via Precision, Recall and F-score, hence providing unbiased evidence to either support or negate $\mathbf{H_1}$ and secondary hypothesis. Additionally, a local off-line MySql based dataset for acquisition of mNGD (2) distances was used to simulate an environment similar to that of a company Intranet where the public domain is not accessible. Using a relatively small (when compared to the WWW) dataset provides further evidence to support or negate $\mathbf{H_{1b}}$ hypothesis within a realistic environment.

4.5 Case study 2: University of Westminster SRS module search

University of Westminster SRS module search system was selected as a subject for this study because it provides a means to evaluate primary and secondary hypothesis (see section 5.3.4 - Study result analyses) in an environment that is very different to the SourceForge.net study. The SRS module search was further able to provide an easily managed data-based personalisation solution.

4.5.1 Research principles

An assortment of subjects with SRS related experience ranging from limited (first year Computer Science undergraduates) to moderate (second year Computer Science undergraduates) was asked to participate. In every case, participation has occurred on a voluntary basis resulting in a total of 51 participants taking part in this study. There are no ethical implications as only statistical data required for Precision, Recall and F-Score calculations were recorded; no record of any personal information (such as age, name, gender, etc.) was made (with the sole exception of participants being in their first or second year of study). None of the participants have had any prior knowledge relating to this study or research. A standard uniform description of a task and of the purpose of this study was provided prior to participation taking place (Appendix B6). Participants had an option to withdraw from the study at any point during their participation. A further option of obtaining study results (once compiled) was made available.

4.5.2 Study participants

To allow for variation and to ensure minimal bias of results a total of 51 subjects from a variety of School of Informatics undergraduate courses participated in the study. These participants were randomly selected (a permission from the

module leader was acquired) for voluntary participation with no preference for age, gender, experience or course. Due to random variation of backgrounds and computer experience, it is possible to generalise that a greater number of participants would not have had any significant impact on results (assuming experimental setup and data are to remain unmodified and participants unbiased). The following is the general description of the two groups of subjects that have participated in this study.

- First year undergraduate students

A total of 22 first year students from the School of Informatics participated in the study. However, 2 queries were found to be identical, thus the total number of first year student queries acquired for the study was 21 (1 was removed). This group of participants had a low degree of familiarity with the task (though, due to general computer background, some familiarity could be assumed). Furthermore, participants were not in any way aware of this research or of the test platform used. Nevertheless, some general familiarity with the process of searching (for example using Google) could be assumed. It should further be noted that, due to being in their first year, this group of students did not have much familiarity with SRS module search facility.

- Second year undergraduate students

A total of 29 second year students from the School of Informatics participated in this study. However, 10 queries were found to be identical and 1 was erroneous, thus the total number of second year student queries acquired for the study was 23 (6 were removed). All students had participated on a voluntary basis. However, unlike with first year students, a degree of familiarity with SRS module search facility has been present. In every other respect this group was treated in the same manner: participants were not in any way aware of this research or of

the test platform used, and no information beyond the standard case study description was provided.

4.5.3 Study environment and platform

For this study a single SemaCS test platform interface was deployed (shown Figure 17). The study was executed on a stand-alone Apache Tomcat 6.0 web server running on Samsung NP-Q1 UMPC. Prior to the study taking place a

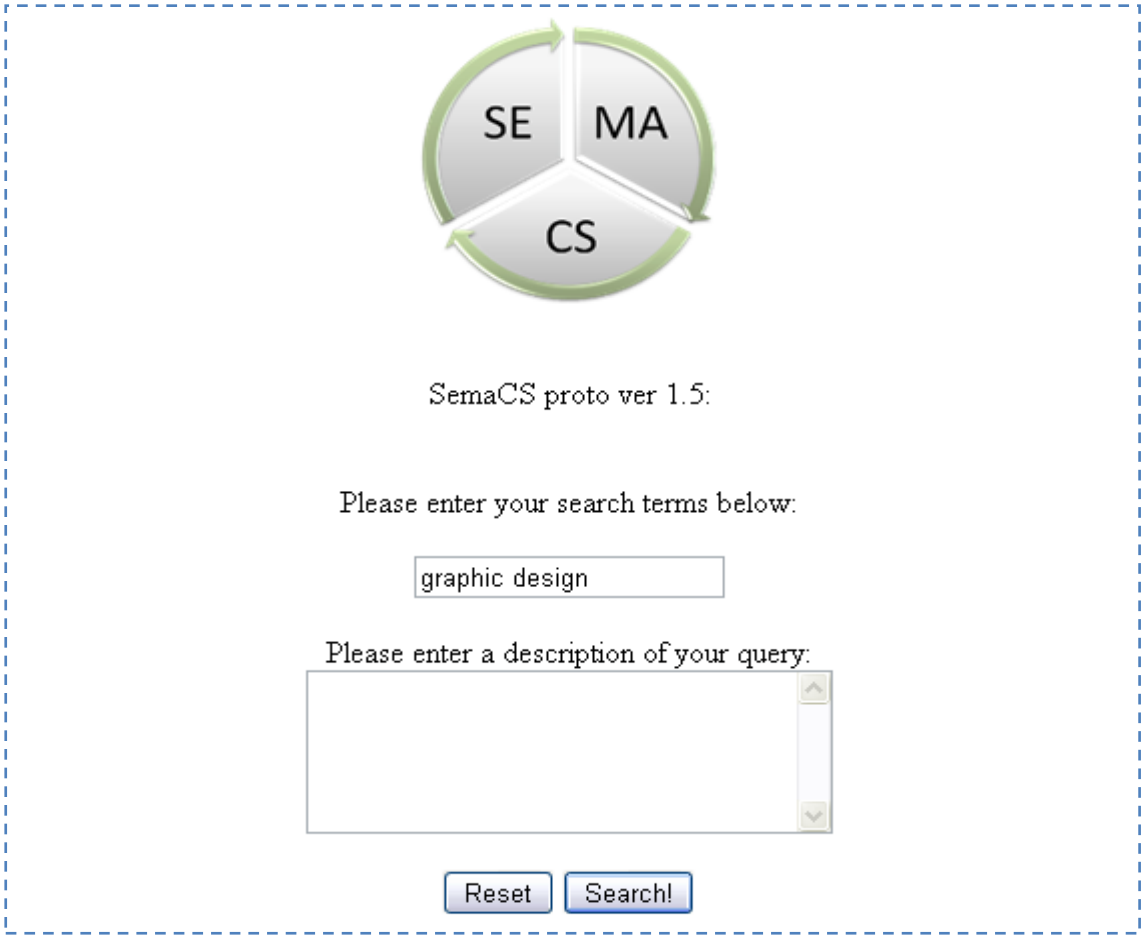

Figure 17: Case study 2 SemaCS search interface

complete set of School of Informatics first year and second year module descriptions were extracted from SRS module search system and corresponding study domain taxonomys were generated (two models for two years). Once study commenced domain taxonomys, environment and test platform have remained unmodified for the entire duration of the case study.

4.5.4 Study data

For this study first and second year University of Westminster School of Informatics modules were used. A total of 69 first year and 82 second year modules were acquired (Appendix B1 and B2). However, unlike the first case study, WWW was used as a source to extract mNGD (2) semantic relation distances from. This was done to emulate an environment similar to that of a company search engine where the public domain is accessible. It should be noted that WWW access was attained by means of Yahoo! interface. Nevertheless, the way WWW was accessed had no significant impact on mNGD (2) because, unlike NGD (1), it is scale invariant.

This study consisted of two parts:

- Part one only considered module titles so as to be able to directly compare against SRS module search facility (which only uses module titles to search)

- Part two, although running within the same environment, further considered module descriptions (hence providing data-based personalisation).

Corresponding domain taxonomies (first and second year with and without module descriptions) were generated. Once generated, domain taxonomies have remained fixed. As a result, the only variables in this study were participant queries used and data-based personalisation.

4.5.5 Study data processing

mNGD (2) was used (see section 3.1.2 - Semantic distance acquisition) to process the data in order to generate indices and domain taxonomy. However, before indexing and model generation took place, case study data was pre-processed using GATE (Cunningham, Maynard et al. 2002) to remove common POS and stop words.

4.5.6 Study data collection procedures

There are no ethical implications as the only statistical data (see Table 5) required for Recall, Precision and F-score calculations was gathered. No record of any personal information was made (with the sole exception of participants being in their first or second year of study). Data was collected automatically via

Collected data	Description
Participant's query	Keywords used as a search string
Time search started	System time search has commenced
Search results	Component Ids and scores
Result return time	System time results were displayed
Participant section	Component Ids (selected from returned results)

Table 5: Case study 2 automatically logged data (definition)

an embedded logging facility with participant query, scenario number and query description (if provided) recorded via the search interface (see Figure 17) and participant selection(s) recorded via the result interface (see Figure 18). Although an option to select a match(es) was made available (via a check box within the results interface), these selection were not made use of for evaluation purposes. This is the case because participants could not be depended on to provide complete, unbiased and accurate information.

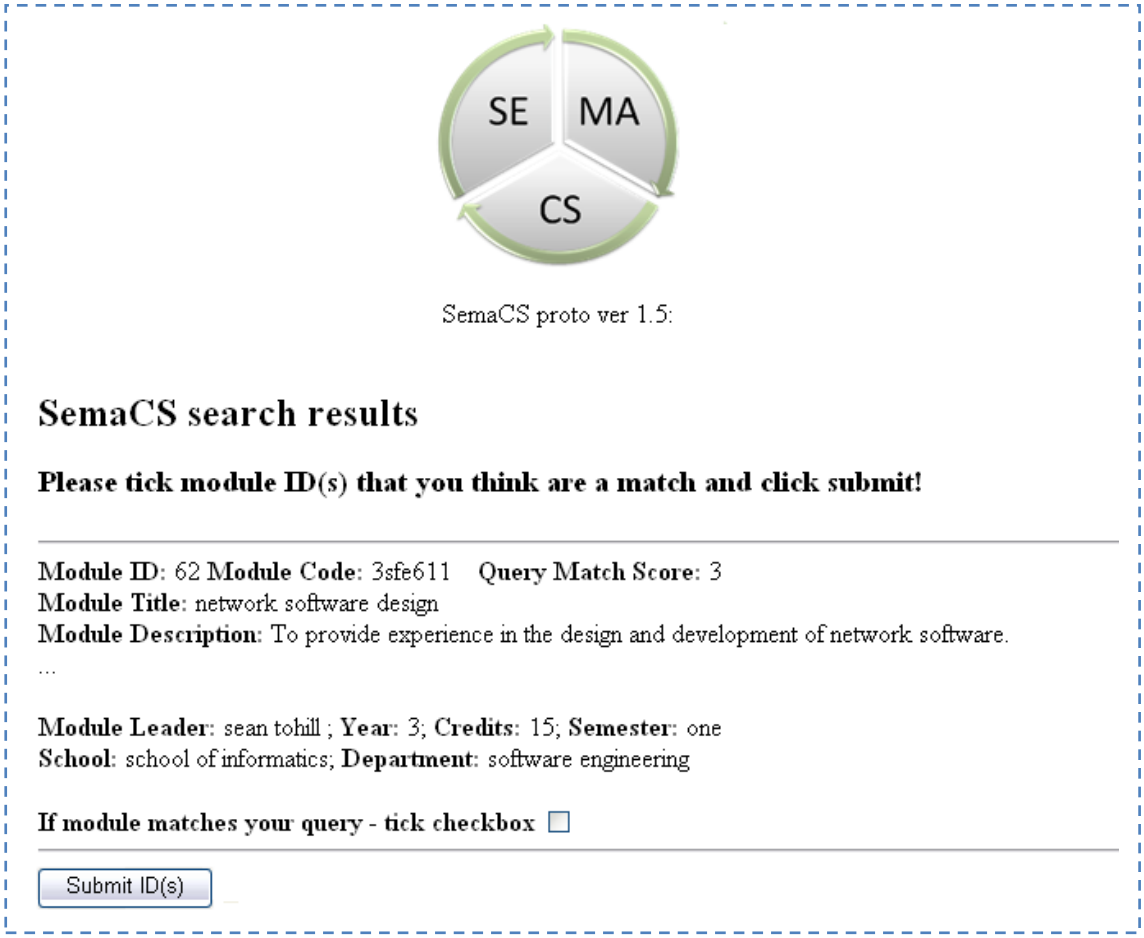

Figure 18: Case study 2 SemaCS results interface

4.5.7 Procedures

All study participants were given the same task to complete – search for free choice modules within the School of Informatics. A uniform explanation of this study as well as a description of the task to be performed was provided to every participant (see Appendix B6) with participants query and results being automatically logged (see Appendix B3). A sample of the log is shown in Figure 19. To perform Precision, Recall, F-score and MRR analyses, experiment log entries were imported into a Microsoft Excel spreadsheet on per query basis (a single line of the log representing a single query). These queries were further split (using semicolons as a separator) into columns representing scenario number, user request (search) string, and a list of matching modules returned by SemaCS. Imported data were then analysed using Precision, Recall, F-score and MRR formulas (see section 4.2 SemaCS evaluation criteria) to generate per query, per Year and total experiment average values shown in Appendix B4 and B5. These data were also used to generate per Year and total experiment 11-point average Precision Recall curves (see section 5.3.3 Combined Years 1 and 2 study result analyses). Upon completion of the data collection, all search

SemaSearch:: Description: graphic design; Request string: graphic design; Time search started: Thu Mar 12 11:39:22 GMT 2009; Search term graphic matches T3: programming; Score: 0.13729591111704326; Search term design matches T3: design; Score: 6.439114970235934E-5; Matched modules: ID: 62 Score: 3| ID: 49 Score: 3| ID: 48 Score: 3| ID: 42 Score: 3| ID: 25 Score: 3| ID: 23 Score: 3| ID: 9 Score: 3| ID: 8 Score: 3| ID: 34 Score: 2| ID: 59 Score: 2| ID: 61 Score: 2| ID: 3 Score: 2| ID: 28 Score: 2| ID: 66 Score: 2| ID: 40 Score: 1| ; Time results returned: Thu Mar 12 11:39:39 GMT 2009; User returned results at: Thu Mar 12 11:40:04 GMT 2009; User section: 40,

Figure 19: University of Westminster SRS case study log sample

queries were also repeated using SRS module search facility (in order to make a direct statistical comparison) and the SemaCS with data-based personalisation (to identify the impact of personalisation on results).

4.5.8 Summary

University of Westminster SRS module search case study was performed on three distinct platforms: SemaCS with and without data-based personalisation and SRS module search facility. Unlike Sourcefore.net case study, access to the public domain was provided (by means of Yahoo!) hence taking validation outside the boundaries of a constrained predictable environment. Another significant difference was the way scenarios were created and assigned: with the Sourcefore.net study there was a fixed predefined number of expert generated scenarios; this was not the case with SRS study where participants were allowed to ‘free search’ for anything of interest within the study domain. These differences provide for a more realistic representation of ‘real’ world requirements and, as a result, a more realistic means of evaluation.

4.6 Conclusion

This chapter has defined the chosen methodology, rationale and methods used to evaluate the primary and secondary hypothesis. It has also described three case studies: Pilot, SourceForge.net and University of Westminster SRS module search. All these studies were designed to provide evidence either supporting or negating $\mathbf{H_1}$ and secondary hypothesis. However, each case study was also different in domain of application as well as deployment environment. It should also be noted that although mNGD (2) was not directly evaluated against other weighting measures because it provides results comparable to NGD (1) (see section 5.1 Pilot study) evaluation performed by the authors on the NGD (1) algorithm equally hold for mNGD (2).

The pilot study has had a positive impact by providing initial evidence supporting $\mathbf{H_1}$ hypothesis as well as the underlying principle approaches (see section 5.1 Pilot study). It has further aided with focusing and designing SourceForge.net and University of Westminster Student Record System module search studies. Furthermore, a number of issues with the SemaCS implementation and design were identified (and corrected) hence avoiding any negative impact on the primary case studies.

SourceForge.net study was performed on two distinct systems: SemaCS and Text-based as a result providing direct unbiased evidence supporting $\mathbf{H_1}$ and secondary hypothesis (see section 5.2 SourceForge.net case study). Additionally, a local off-line MySql based dataset for acquisition of mNGD (2) distances was used to simulate an environment similar to that of a company Intranet where public domain is not accessible. Using a relatively small (when compared to the WWW) dataset provided further evidence in support of $\mathbf{H_{1b}}$ hypothesis (mNGD) within a realistically modelled environment.

University of Westminster SRS module search study was performed on three distinct platforms: SemaCS with and without data-based personalisation and SRS module search facility. Unlike SourceForge.net case study, access to the public domain was provided (by means of Yahoo!) hence taking validation outside the boundaries of a constrained predictable environment. Additionally, SourceForge.net study was not subject to a fixed number of predefined expert generated scenarios; study participants were allowed to 'free search' for anything of interest within the study domain hence providing further evidence in support of $\mathbf{H_1}$ and secondary hypothesis in a realistically modelled representation of 'real' world requirements and application (see section 5.3 University of Westminster case study). This study has also been used to provide evidence supporting of $\mathbf{H_{1d}}$ hypothesis (data-based personalisation).

Chapter 5 Result analyses

In chapter 3 the 3 stages of SemaCS algorithm: text pre-processing and semantic distance acquisition, domain taxonomy generation and textual description index generation, search and personalisation were defined. Chapter 3 has further defined the SemaCS prototype created to provide a means of evaluating the primary and secondary hypotheses and elaborated on design and implementation decisions. In chapter 4, the chosen evaluation criteria of Precision, Recall and F-score were introduced. Chapter 4 also further defined the methodology, rationale and case studies used to evaluate the primary and secondary hypothesis.

This chapter presents and analyses the significance of results collected by means of a feasibility study and two case studies described in the previous chapter. The remainder of this chapter is structured in the following way: section 5.1 presents and analyses feasibility study results; section 5.2 presents and analyses case study 1: SourceForge.net results; and finally, section 5.3 presents and analyses case study 2: University of Westminster SRS module search results.

5.1 Pilot study

The ability to detect a degree of relatedness between words is a key feature of the SemaCS approach. This study was designed (see section 4.3 - Pilot study) to evaluate the feasibility and impact of a proposed modification to NGD (1) – mNGD (2) (see section 3.1.2 - Semantic distance acquisition) on that ability. Thus, providing initial evidence either supporting or disproving hypothesis $\mathbf{H_1}$.

To evaluate the suitability and performance of mNGD (2) and NGD (1) algorithms two iterations of the same experiment (see section 4.3 - Pilot study) were implemented: first by means of original NGD (1) algorithm and second by means of mNGD (2) algorithm. Using these algorithms, semantic distance relation values were calculated for a set of 31 word pairs acquired from Rubenstein and Goodenough 1965 study. Expert generated relatedness scores for the above set of 31 word pairs were also acquired from the same study; the data formed a golden standard against which mNGD (2) and NGD (1) results were evaluated. Thus providing evidence with view to support hypothesis $\mathbf{H_{1b}}$. Furthermore, due to experimental data not being from the software component domain, evidence was also to be provided in support of hypothesis $\mathbf{H_{1a}}$.

5.1.1 mNGD - NGD experiment

Although both experiments were performed within identical environments and using identical datasets each implemented a different algorithm. This section elaborates on NGD (1) based experiment results (see Table 6) which demonstrated a degree of correlation between NGD (1) obtained scores and the gold standard (expert generated) scores. Consequently, the data provides positive evidence supporting NGD (1) algorithm applicability for detection of semantic relatedness. Although, when interpreting these results, it should be noted the gold standard and NGD (1) scoring systems are of different scale: golden standard scores are rated from 0 to 4 (with a maximum relation value represented as 4) while NGD (1) scores are rated from infinity to 0 (with a maximum relation value represented as 0). As could be seen from Table 6, NGD (1) was able to detect an existence of a relationship for all given word pairs. However, its performance in detecting the significance of that relationship appeared to be deficient (e.g., two synonymous words ‘automobile’ and ‘car’ were scored at 1.6269201 – a weak relation).

Word1	Word2	Original Score	NGD Score	mNGD Score
cord	smile	0.02	1.5156207	0.6620554
rooster	voyage	0.04	1.0472801	0.6207060
noon	string	0.04	1.2316870	0.5216656
fruit	furnace	0.05	0.9888798	0.5432781
autograph	shore	0.06	0.9920439	0.5656396
automobile	wizard	0.11	1.4345983	0.4896297
mound	stove	0.14	0.8542504	0.4732687
grin	implement	0.18	1.0968248	0.5297539
asylum	fruit	0.19	1.0828027	0.5548039
graveyard	madhouse	0.42	0.6202633	0.4513386
glass	magician	0.44	1.1675092	0.7267373
boy	rooster	0.44	1.0825422	0.6495127
cushion	jewel	0.45	0.9876309	0.4865527
monk	slave	0.57	0.9027979	0.4107694
asylum	cemetery	0.79	0.9300117	0.4549333
cord	string	3.41	1.4487783	0.6259341
glass	tumbler	3.45	1.0165516	0.8569555
grin	smile	3.46	1.3986148	0.7626095
serf	slave	3.46	0.6427643	0.6434726
journey	voyage	3.58	1.2688352	0.4124394
autograph	signature	3.59	1.2451353	0.8189561
coast	shore	3.6	1.2399979	0.4396268
forest	woodland	3.65	1.2286520	0.6414237
implement	tool	3.66	0.9959402	0.3514267
cock	rooster	3.68	1.1774510	0.7567024
boy	lad	3.82	1.1165463	0.5949166
cushion	pillow	3.84	0.9078467	0.4347350
cemetery	graveyard	3.88	0.7771304	0.3976809
automobile	car	3.92	1.6269201	0.5138451
midday	noon	3.94	0.7512109	0.4903489
gem	jewel	3.94	1.1562524	0.4589368

Table 6: Pilot study experiment results

Such a low accuracy of relatedness strength detection was unexpected, hence prompting validation of NGD (1) generated results. Experiment results were manually verified to ensure absence of any implementation errors. Verification was performed for every word pair shown in Table 6; a manual Google query was formed and the NGD (1) relation scores manually calculated. Although, for simplicity, manual calculations were performed to 5 decimal places (not 8 as with SemaCS) and the same results were found. Hence the accuracy of SemaCS based NGD (1) implementation was verified.

In order to verify the correctness of the SemaCS algorithmic interpretation of NGD (1) algorithm, data for “horse” and “rider” were acquired from the original NGD publication (Cilibrasi and Vitanyi 2004) and used to calculate a semantic distance between these terms. Using these data (“horse” hits = 46700000, “rider” hits = 12200000, “horse” “rider” hits = 2630000, N = 8058044651) calculations were performed manually and using SemaCS with both resulting in an identical score for NGD horse rider ≈ 0.44305631 which also correlated with results published in (Cilibrasi and Vitanyi 2004) – NGD horse rider ≈ 0.443. However, using live data obtained from Google has provided a different score (NGD horse rider ≈ 1.04774933). The only feasible explanation for such a significant deviation appears to be a considerable change in N (the number of pages referenced by Google) since the data were published. For this reason, a comparison of the two scores cannot be made as they are of different ‘magnitude’.

With SemaCS algorithmic interpretation and implementation of NGD (1) verified – based on analysis of Table 6 results and reasoning – it was concluded that NGD (1) does not provide a scale-invariant solution. Consequently, it was also concluded that NGD (1) cannot ensure result consistency over a period of time as semantic relation significance (score) varies significantly over time due to

being dependent on **N** value (which can only be estimated). Furthermore, experiment results also demonstrated that mNGD (2) is capable of providing a comparable solution of semantic relation acquisition not dependant on **N**; therefore, providing substantial evidence supporting hypothesis $\mathbf{H_{1b}}$. Moreover, as experiment word pairs are from a number of unrelated domains, initial evidence supporting hypothesis $\mathbf{H_{1a}}$ was also provided.

5.1.3 Study result analyses

To evaluate the mNGD (2) ability to detect the strength of a relationship to that of NGD (1) and compare the two approaches to the gold standard, a means of statistical comparison had to be found. Although mNGD (2) could be easily compared to NGD (1) (their weighting approaches i.e. score systems are identical), it could not be compared to the expert generated gold standard (which implements an entirely different scoring system). Consequently, to allow for a meaningful comparison to take place, golden standard relation scores were converted to a compatible form.

As neither mNGD (2) nor NGD (1) generated relation scores of significance less than 2 (see Table 6) a conversion scale of 2 was assumed. Golden standard scores (weighted 0 to 4, with 4 representing maximum strength of relation) were inverted by subtracting them from 4. This translated scores to a form were, similarly to mNGD (2) and NGD (1), 0 represented the maximum strength of relation. These inverted scores were then divided by 2 to convert them to a compatible scale of magnitude (changing the scale of representation does not modify score significance in statistical terms, for example, 2 of 4 is the same in magnitude weight as 4 of 8). With all relation weights residing within identical scale of representation a meaningful comparison could be made. Figure 13 represents the result of this comparison – a relation of mNGD (2), NGD (1) and golden standard scores.

As shown in Figure 20 (when contrasted against the golden standard) mNGD (2) is 50% better than NGD (1) at detecting stronger semantic relations while NGD (1) is 50% better than mNGD (2) at detecting weaker semantic relations. This association between the two algorithms as well as the fact that line plots (apart for magnitude of scale) appear to be very similar, prompted further investigation. However, at this point it was concluded that hypothesis $\mathbf{H_{1b}}$ is validated as mNGD (2) was able to generate semantic distance relations comparable to NGD (1) without relying on **N**.

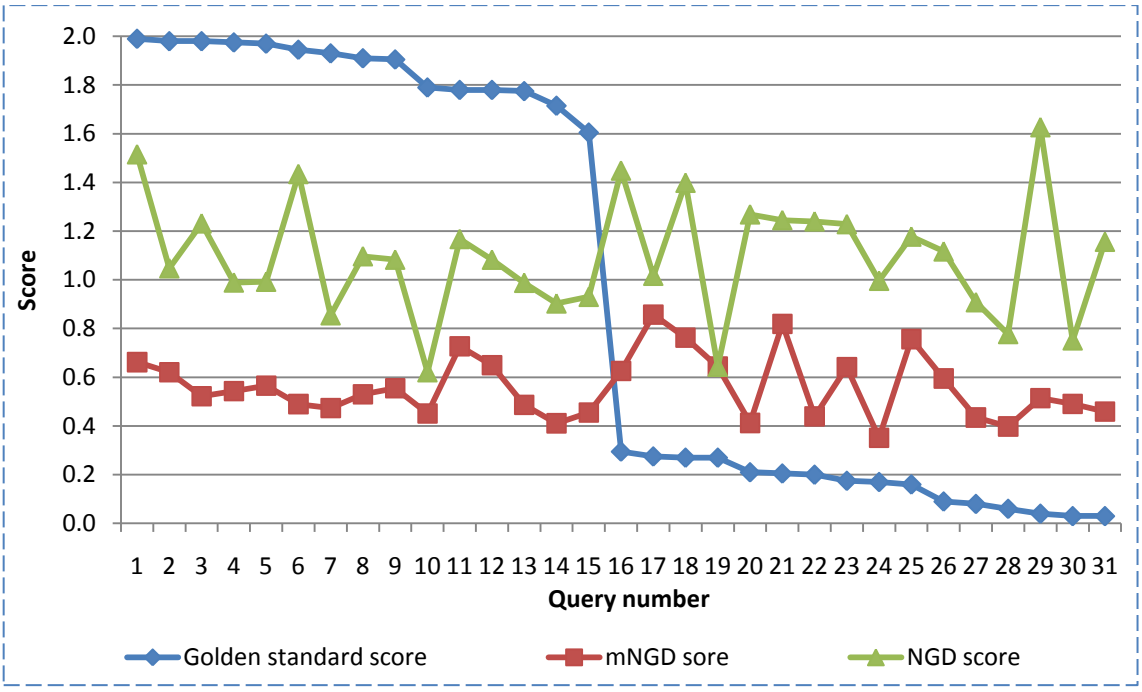

Figure 20: Golden standard, mNGD and NGD relation scores

Due to an observed association between mNGD (2) and NGD (1) scores the magnitude of NGD (1) based scores was investigated further. As a result of this investigation it was proposed that the difference in NGD (1) scale magnitude is likely to be caused by differences between actual and SemaCS representations of the **N** value. However, as **N** can only be estimated (Kunder 2009), a different

approach was used to better compare the two algorithms. Similarly to the gold standard, a scale magnitude conversion process for NGD (1) scores was defined. However, it should be noted that, in this case, changing magnitude of scale did modify the score significance. This is acceptable because, instead of direct comparison of score significance, an association between the two algorithms was being considered. To make such an association explicit both algorithms must coexist within the same plane. The conversion process was performed by calculating an average variance difference between NGD (1) and mNGD (2) obtained scores (average variance = 0.53524889) and then subtracting it from NGD (1) generated scores. Figure 21 demonstrates the result of this comparison between converted scores.

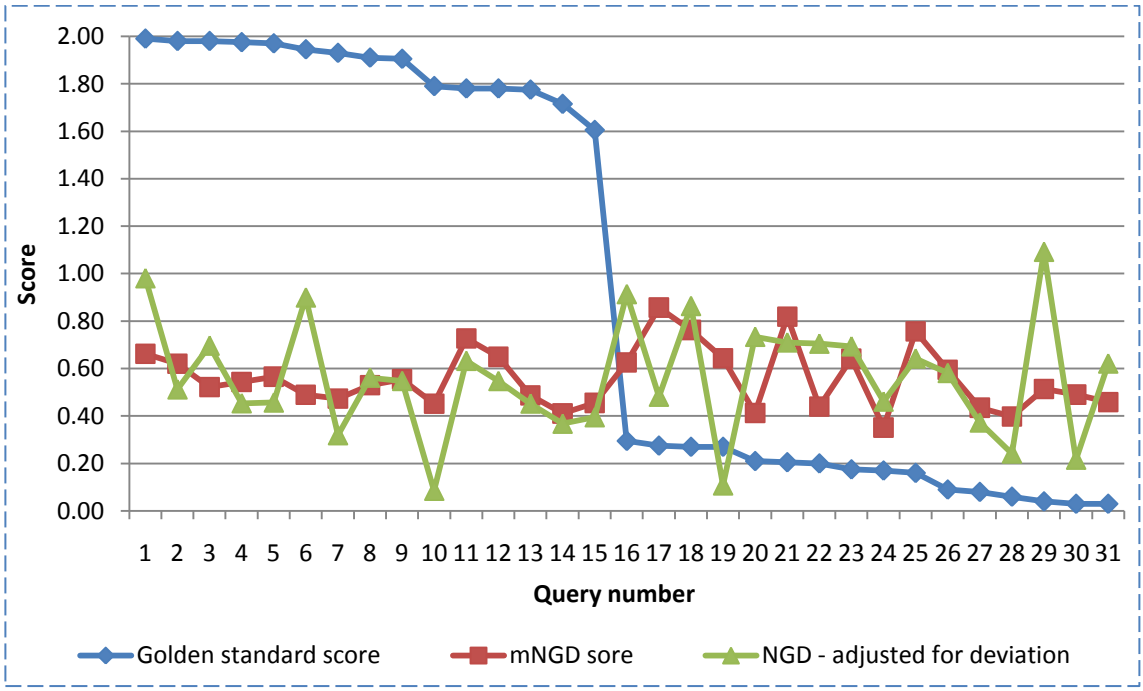

Figure 21: Golden standard, mNGD and converted NGD relation scores

As demonstrated in Figure 21, NGD (1) and mNGD (2) generated relation scores are comparable. Although this could imply that (given a 'true'

representation of **N** could be acquired) mNGD (2) scores would be identical to NGD (1), empirical evidence to support this hypothesis is not available. While being a curious consequence that warrants further investigation, it is not perceived as a threat to feasibility of mNGD (2) application to support $\mathbf{H_1}$ hypothesis because (regardless of mNGD (2) relation scores being identical to NGD (1) or not), mNGD (2) provides a comparable but **N** independent solution.

5.1.4 Summary

The pilot study has provided a means to evaluate the feasibility and impact of mNGD (2) on SemaCS ability to detect a degree of relatedness between words. It has further provided positive evidence supporting $\mathbf{H_1}$ hypothesis by providing evidence supporting mNGD (2) ability to discover semantic distances (hypothesis $\mathbf{H_{1b}}$). These experiments have also provided further rationale supporting mNGD (2) modification: unmodified NGD (1) cannot provide a scale invariant solution and, consequently, cannot ensure result consistency over a period of time. Additionally, the golden standard used for result evaluation purposes did not belong to the software component domain which provided positive evidence in support of hypothesis $\mathbf{H_{1a}}$. This study has also identified a close association between mNGD (2) and NGD (1) generated scores. However, due to its low (or nonexistent) impact on $\mathbf{H_1}$ hypothesis and resource limitations this association was not investigated further.

5.2 SourceForge.net case study

SourceForge.net was chosen as a subject for a study due to being a large and well known software portal as well as its use in a number of related evaluations (see section 4.4 - Case study 1: SourceForge.net). Similarly to the pilot study a single experiment was implemented (see section 4.4 - Case Study 1: Sourceforge.net). This experiment was repeated on two platforms: SemaCS-based and Keyword-based. Using two distinct platforms provided for a better means of evaluation with all procedures designed in such a way as to facilitate statistical evaluation and comparison via Precision, Recall, F-score and MRR hence providing evidence to either support or negate $\mathbf{H_1}$. Additionally, by being implemented within software component domain (unlike the pilot and second case studies), this study provided further evidence supporting hypothesis $\mathbf{H_{1a}}$. Furthermore, due to an Intranet based local dataset used for mNGD (2) semantic distance relation acquisition, this study provided positive evidence supporting $\mathbf{H_{1b}}$. Finally, because a domain taxonomy was automatically generated as part of this study, evidence supporting $\mathbf{H_{1c}}$ was provided.

5.2.1 SemaCS experiment

For this experiment SourceForge.net software component database was used as a source to extract mNGD (2) semantic relation distances from (see section 4.4.4 - Study Data). This was done to simulate an environment similar to that of a company Intranet where the public domain is not accessible. Using a relatively small dataset (when compared to the WWW) provided further evidence to support $\mathbf{H_{1b}}$ hypothesis within a realistic environment. Experiment procedures and data structures are described in section 4.4, resulting logs and calculation tables can be found in Appendix A2 and A3. However, a concise version of the results is given in Table 7. It should be noted that SemaCS results are presented using a score cut-off point (see section 4.2.1 - Precision) where matches with a score less than or equal to 3 were removed, unless only one

Q №	Scenario	Search query	Precision	Recall	F-score	RR
2	1	download file server	0.250000	1.000000	0.400000	0.00
3	1	ftp upload software	0.428571	1.000000	0.600000	1.00
21	1	software to upload files	0.200000	0.666667	0.307692	0.50
36	1	application dowload upload server	0.176471	1.000000	0.300001	0.25
41	1	file upload	0.000000	0.000000	0.000000	0.00
56	1	download onto server	0.214286	1.000000	0.352942	0.00
61	1	ftp client	0.666667	0.666667	0.666667	0.50
76	1	ssh secure shell	0.000000	0.000000	0.000000	0.00
5	2	network traffic monitoring tools	0.166667	1.000000	0.285715	1.00
8	2	network traffic monitor	0.166667	1.000000	0.285715	1.00
22	2	network traffic monitor tool	0.142857	1.000000	0.250000	0.33
37	2	monitor network traffic	0.166667	1.000000	0.285715	0.50
42	2	network traffic	0.500000	1.000000	0.666667	0.50
57	2	netwrok traffic	0.000000	0.000000	0.000000	0.00
62	2	monitor network traffic	0.166667	1.000000	0.285715	0.50
77	2	monitor network traffic	0.166667	1.000000	0.285715	0.50
9	3	software download free remote PC control	0.083333	1.000000	0.153846	0.50
12	3	remote application to control computer	0.200000	1.000000	0.333333	0.25
23	3	remote control	0.333333	1.000000	0.500000	1.00
38	3	application for remotly control computer	0.000000	0.000000	0.000000	0.00
43	3	remote desktopping	1.000000	1.000000	1.000000	1.00
58	3	remote control computer	1.000000	1.000000	1.000000	1.00
63	3	controling computer	0.000000	0.000000	0.000000	0.00
78	3	remotly control computer	0.000000	0.000000	0.000000	0.00
13	4	video audio player	1.000000	0.800000	0.888889	1.00
16	4	mpeg4	0.333333	0.200000	0.250000	0.50
24	4	audio	1.000000	0.400000	0.571429	1.00
39	4	media player	1.000000	0.200000	0.333333	1.00
44	4	mpeg player	1.000000	0.200000	0.333333	1.00
59	4	media applications	0.000000	0.000000	0.000000	0.00
64	4	video playback and audio playback	0.833333	1.000000	0.909091	1.00
79	4	media player to plsy video/audio	0.166667	0.200000	0.181818	0.00
17	5	sound conversion	0.333333	1.000000	0.500000	1.00
20	5	sound file converter	0.500000	1.000000	0.666667	1.00
25	5	sound file conversion	0.500000	1.000000	0.666667	1.00
40	5	sound convertor	0.076923	1.000000	0.142857	0.00
45	5	mp3 maker	0.200000	1.000000	0.333333	0.50
60	5	modify sound files	0.000000	0.000000	0.000000	0.00
65	5	mp3 conversastion	0.500000	1.000000	0.666667	0.50
80	5	sound file modifier and converter	0.052632	1.000000	0.100001	0.00

Table 7: Case study 1 SemaCS-based experiment results

item were to remain, in which case matches with a score less than or equal to 2 were removed. This cut-off rule was generated to provide for an optimal level of Precision and Recall. 40 automatically logged entries resulting from this experiment (given in Table 7) were analysed via an Excel spreadsheet (see Appendix A3 A) to arrive at a per scenario and experiment F-score, MRR and average Precision and Recall values shown in Table 8.

	Precision	Recall	F-Score	MRR
Scenario 1 average	24.20%	66.67%	35.51%	0.28
Scenario 2 average	18.45%	87.50%	30.48%	0.54
Scenario 3 average	32.71%	62.50%	42.94%	0.47
Scenario 4 average	67.08%	37.50%	48.11%	0.69
Scenario 5 average	27.04%	87.50%	41.31%	0.50
Experiment Total	**33.90%**	**68.33%**	**45.31%**	**0.50**

Table 8: Case study 1 SemaCS-based experiment results (average)

Although average Recall is only ≈ 68% and MRR is only 0.50 (or second element of the result set), it should be noted that these results and calculations were drawn from complex data and real queries. Furthermore, SemaCS has dealt with imprecise and error prone data and queries (both logically as well syntactically) and has received no expert generated training nor did it employ conventional NLP or AI algorithms. Even though higher levels of accuracy were initially expected, the fact that an average of 68.33% Recall and 45.31% F-score was achieved provides positive evidence supporting $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$ and $\mathbf{H_{1c}}$ hypothesis and consequently $\mathbf{H_1}$ hypothesis.

5.2.2 Keyword-based experiment

As with SemaCS experiment data and queries (shown in Table 9) have remained unchanged. However, query results were generated by a simple

Q №	Scenario	Search query	Precision	Recall	F-score	RR
1	1	download file server	0.230769	1.000000	0.375000	1.00
4	1	ftp upload software	0.428571	1.000000	0.600000	1.00
26	1	software to upload files	0.050000	0.666667	0.093023	1.00
31	1	application dowload upload server	0.142857	0.666667	0.235294	0.00
46	1	file upload	0.333333	0.666667	0.444444	0.50
51	1	download onto server	0.250000	0.666667	0.363636	0.50
66	1	ftp client	0.400000	0.666667	0.500000	1.00
71	1	ssh secure shell	0.000000	0.000000	0.000000	0.00
6	2	network traffic monitoring tools	0.200000	1.000000	0.333333	1.00
7	2	network traffic monitor	0.333333	1.000000	0.500000	1.00
27	2	network traffic monitor tool	0.111111	1.000000	0.200000	1.00
32	2	monitor network traffic	0.333333	1.000000	0.500000	1.00
47	2	network traffic	0.333333	1.000000	0.500000	1.00
52	2	netwrok traffic	0.000000	0.000000	0.000000	0.00
67	2	monitor network traffic	0.333333	1.000000	0.500000	1.00
72	2	monitor network traffic	0.333333	1.000000	0.500000	1.00
10	3	software download free remote PC control	0.062500	1.000000	0.117647	1.00
11	3	remote application to control computer	0.024390	1.000000	0.047619	0.50
28	3	remote control	0.166667	1.000000	0.285714	1.00
33	3	application for remotly control computer	0.000000	0.000000	0.000000	0.00
48	3	remote desktopping	0.500000	1.000000	0.666667	1.00
53	3	remote control computer	0.142857	1.000000	0.250000	1.00
68	3	controling computer	0.000000	0.000000	0.000000	0.00
73	3	remotly control computer	0.000000	0.000000	0.000000	0.00
14	4	video audio player	0.500000	0.800000	0.615385	1.00
15	4	mpeg4	0.000000	0.000000	0.000000	0.00
29	4	audio	1.000000	0.800000	0.888889	1.00
34	4	media player	0.500000	0.400000	0.444444	1.00
49	4	mpeg player	0.600000	0.600000	0.600000	1.00
54	4	media applications	0.333333	0.200000	0.250000	1.00
69	4	video playback and audio playback	0.142857	1.000000	0.250000	1.00
74	4	media player to plsy video/audio	0.097561	0.800000	0.173913	1.00
18	5	sound conversion	0.000000	0.000000	0.000000	0.00
19	5	sound file converter	0.000000	0.000000	0.000000	0.00
30	5	sound file conversion	0.000000	0.000000	0.000000	0.00
35	5	sound convertor	0.000000	0.000000	0.000000	0.00
50	5	mp3 maker	0.500000	1.000000	0.666667	1.00
55	5	modify sound files	0.000000	0.000000	0.000000	0.00
70	5	mp3 conversastion	0.500000	1.000000	0.666667	1.00
75	5	sound file modifier and converter	0.029412	1.000000	0.057143	0.00

Table 9: Case study 1 Text-based experiment results

keyword match algorithm described in section 3.3.2 Text-based search. This algorithm searches for an occurrence of each keyword supplied as part of the query in turn; if a match is found it is displayed (a greater number of matches produces a higher score). It should also be noted that both SemaCS and Text-based implementations were restricted to the use of software component descriptions only (any reference to the 'name' of the software component has been removed in all instances). Experimental procedures and data structures are described in section 4.4, resulting logs and calculation tables can be found in Appendix A2 and A3. However, a concise version of the results is given in Table 9. 40 automatically logged entries resulting from this experiment (given in Table 9) were analysed via an Excel spreadsheet (see Appendix A3 B) to arrive at a per scenario and experiment F-score, MRR and average Precision and Recall values shown in Table 10.

	Precision	Recall	F-Score	MRR
Scenario 1 average	22.94%	66.67%	34.14%	0.63
Scenario 2 average	24.72%	87.50%	38.55%	0.88
Scenario 3 average	11.21%	62.50%	19.00%	0.56
Scenario 4 average	39.67%	57.50%	46.95%	0.88
Scenario 5 average	12.87%	37.50%	19.16%	0.25
Experiment Total	**22.28%**	**62.33%**	**32.83%**	**0.64**

Table 10: Case study 1 Text-based experiment results (average)

Although a low F-Score score was achieved, it should be noted that Recall and relevancy levels were high. Due to such high levels of Recall, relevancy and simplicity of implementation Keyword-based approaches are still widely used. However, these approaches are inflexible and not capable of detecting any kind of semantic significance or handle syntactical mistakes (see section 2.3.2 - Intelligent search).

5.2.3 Study result analyses

This study was designed to collect evidence with the view to support primary and secondary hypothesis. For comparison purposes the study experiment was repeated on two distinct platforms: SemaCS and Text-based. Using a per-query F-Score based comparison (shown in Figure 22); SemaCS has clearly outperformed a Keyword-based approach by achieving an average improvement of: 11.61% in Precision, 6% in Recall and 12.49% in F-Score. These improvements provide positive evidence in support of $\mathbf{H_1}$, $\mathbf{H_{1b}}$ and $\mathbf{H_{1c}}$ hypothesis as well as in support of $\mathbf{H_{1a}}$ hypothesis, although evidence supporting $\mathbf{H_{1a}}$ remains partial until combined with results obtained via the second study.

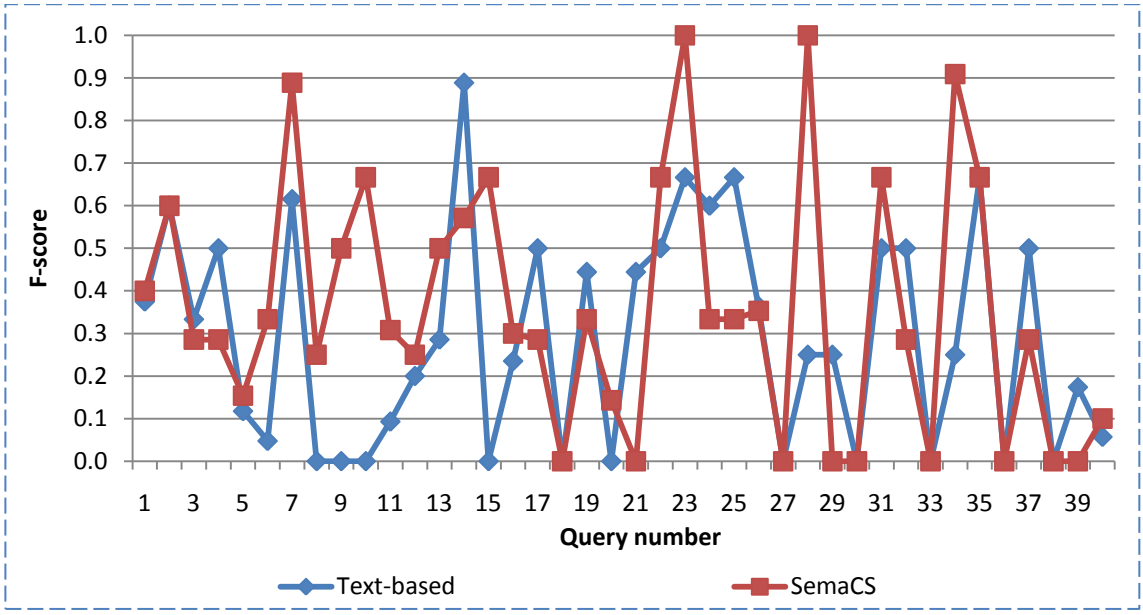

Figure 22: Sourseforge.net case study F-score comparison

However, Precession, Recall and F-score are set-based evaluation measures; they do not take into account the fact that results are sorted in order of relevancy. MRR and interpolated average Precision Recall were selected to

assess SemaCS result relevancy algorithms (see section 4.2 SemaCS evaluation criteria). Following sections make a per scenario comparison between the two platforms:

- Scenario 1

SemaCS has outperformed a Text-based implementation by an average of 1.26% in Precision and 1.37% in F-score with Recall being an identical 66.67%. However as made evident by the MRR values, Text-based implementation has outperformed SemaCS by 0.34. Although SemaCS demonstrated a better overall ratio of relevant to irrelevant result elements when the entire result set is considered, Text-based implementation was better at giving relevant results a higher rank. This is further made evident by the 11-point interpolated average Precision Recall curve shown in Figure 23:

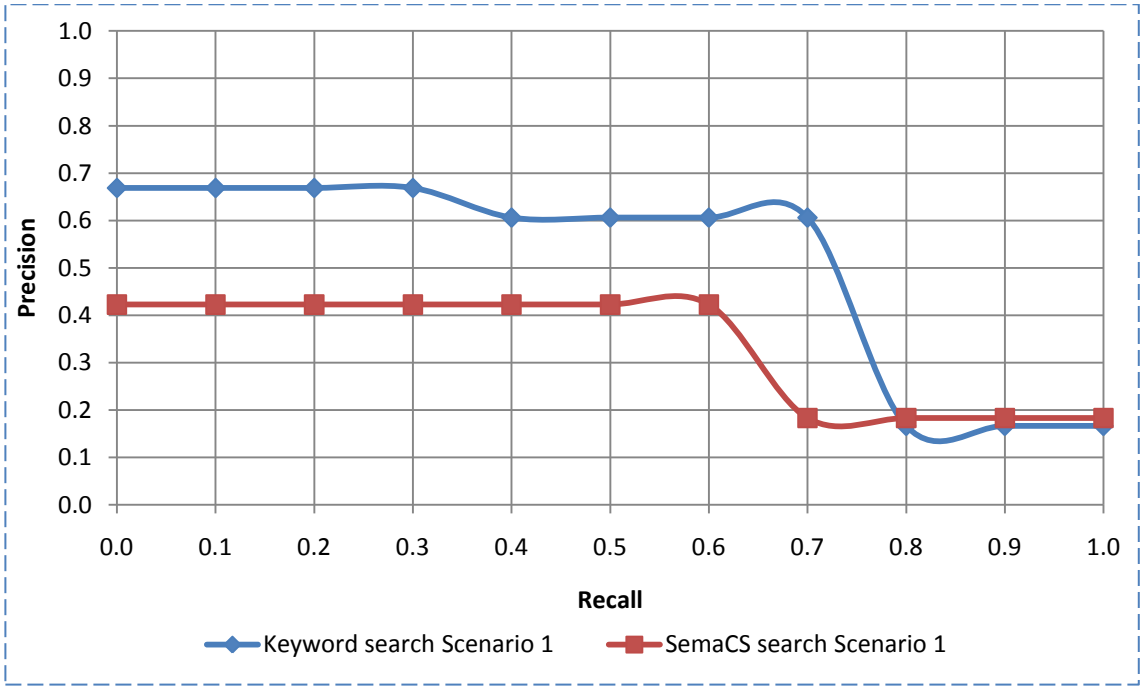

Figure 23: Scenario 1, 11-point interpolated average Precision Recall curves

- Scenario 2

Both implementations achieved 87.5% Recall. However, SemaCS was outperformed by the Text-based implementation at an average of 6.27% in Precision and 8.07% in F-score. SemaCS was further outperformed by 0.33 in MRR. Similarly to scenario 1, Text-based implementation was better at giving relevant results a higher rank, this is further made evident by the 11-point interpolated average Precision Recall curve shown in Figure 24. Although, it should be noted that SemaCS has achieved a MRR of 0.54 (just over every second element), which is a significant improvement of scenario 1.

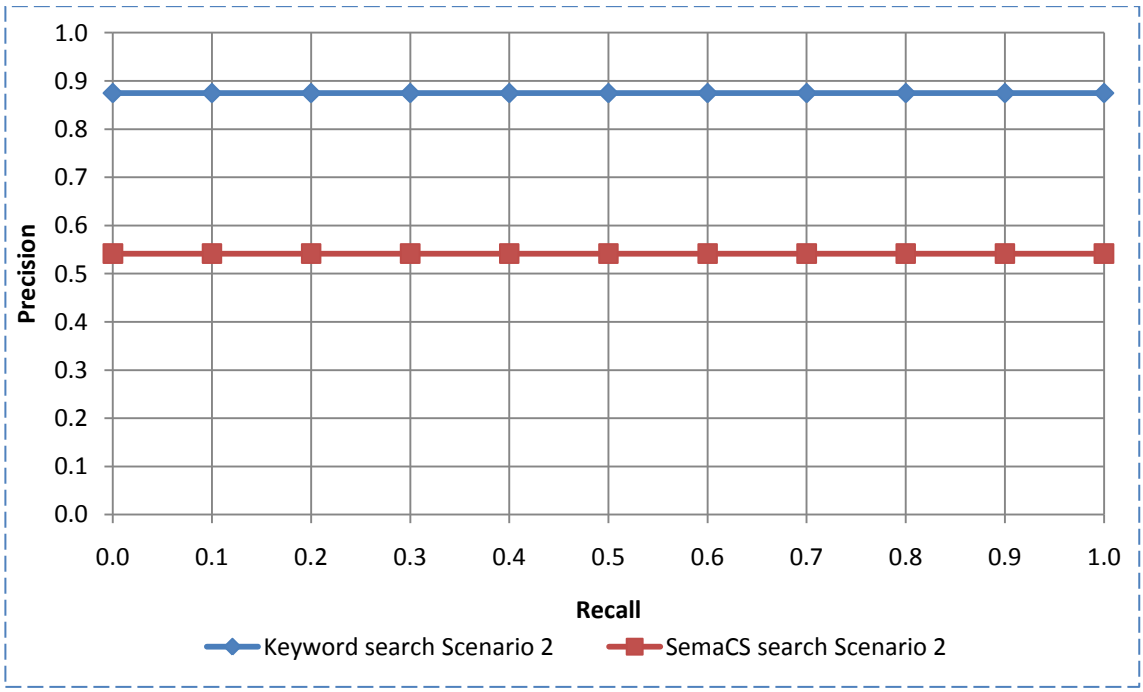

Figure 24: Scenario 2, 11-point interpolated average Precision Recall curves

- Scenario 3

Similarly to scenario 1 SemaCS has outperformed the Text-based implementation by an average of 21.5% in Precision and 23.94% in F-score.

Also similarly to the two previous cases, both implementations have achieved an identical average of 62.5% in Recall. However as made evident by the MRR values, Text-based implementation has outperformed SemaCS by a margin of 0.09. Although SemaCS has demonstrated a better overall ratio of relevant to irrelevant elements when the entire result set is considered, Nonetheless, Text-based implementation was marginally better at giving the relevant elements a higher rank. This is further made evident by the 11-point interpolated average Precision Recall curve shown in Figure 25:

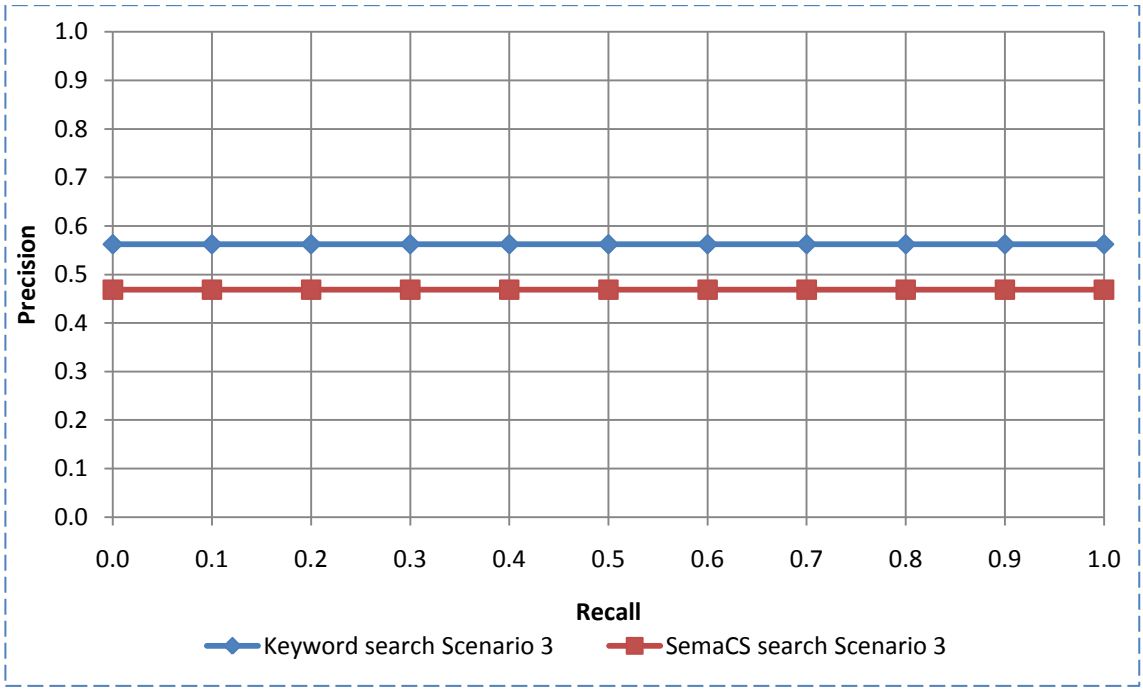

Figure 25: Scenario 3, 11-point interpolated average Precision Recall curves

- Scenario 4

SemaCS has achieved an average improvement of 27.41% in Precision and 1.16% in F-score. However, Text-based implementation has achieved a noticeable improvement of 20% in Recall and 0.19% in MRR. Similarly to previous scenarios, the Text-based implementation has demonstrated that it is

capable of scoring correct matches higher. Although SemaCS was not able to detect a similar number of relevant results, it has managed to return smaller sized result sets as made evident by the overall Precision score. Nevertheless, SemaCS was outperformed in every other respect; this is further made evident by the 11-point interpolated average Precision Recall curves shown in Figure 26:

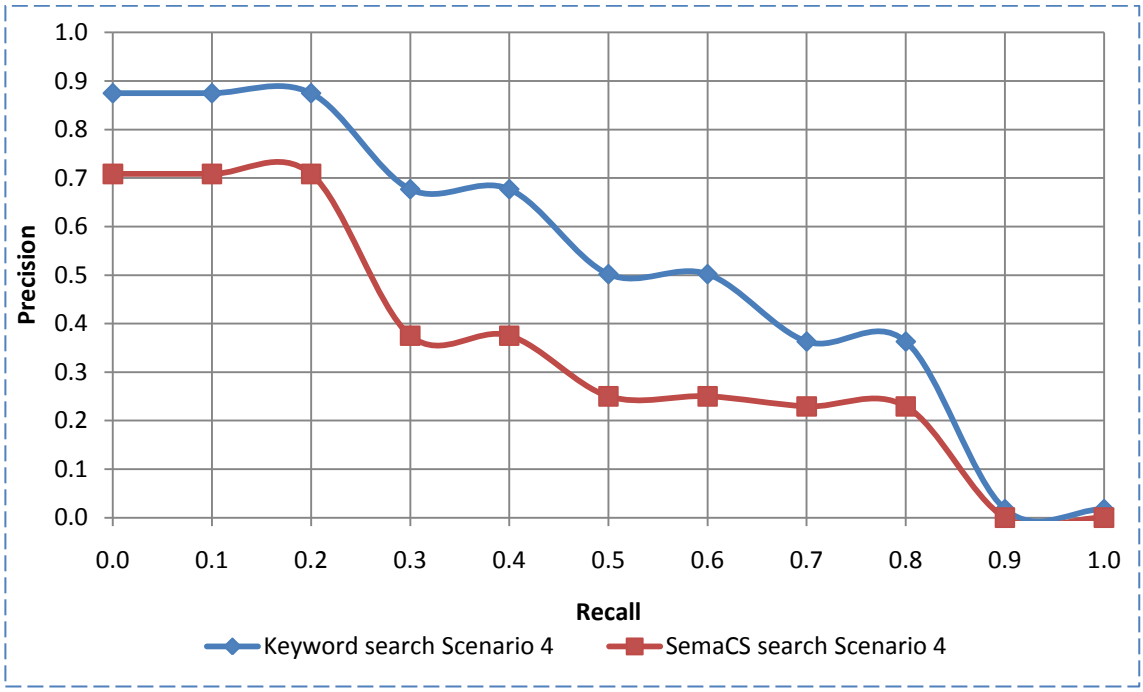

Figure 26: Scenario 4, 11-point interpolated average Precision Recall curves

It should be noted that Scenario 4 is the only scenario that consists of 5 possible matches. Thus the curve representing this scenario is the most descriptive out of the set. Furthermore, due to the increased number of possible matches, unlike with previous scenarios, average Recall scores achieved by the two implementations differ significantly.

- Scenario 5

SemaCS has achieved an average improvement of 14.17% in Precision, 50% in Recall and 22.15% in F-score. This is also the only scenario where SemaCS has outperformed the Text-based implementation by achieving an average improvement of 0.25 in MRR. This is further reflected on the 11-point interpolated average Precision Recall curve shown in Figure 27:

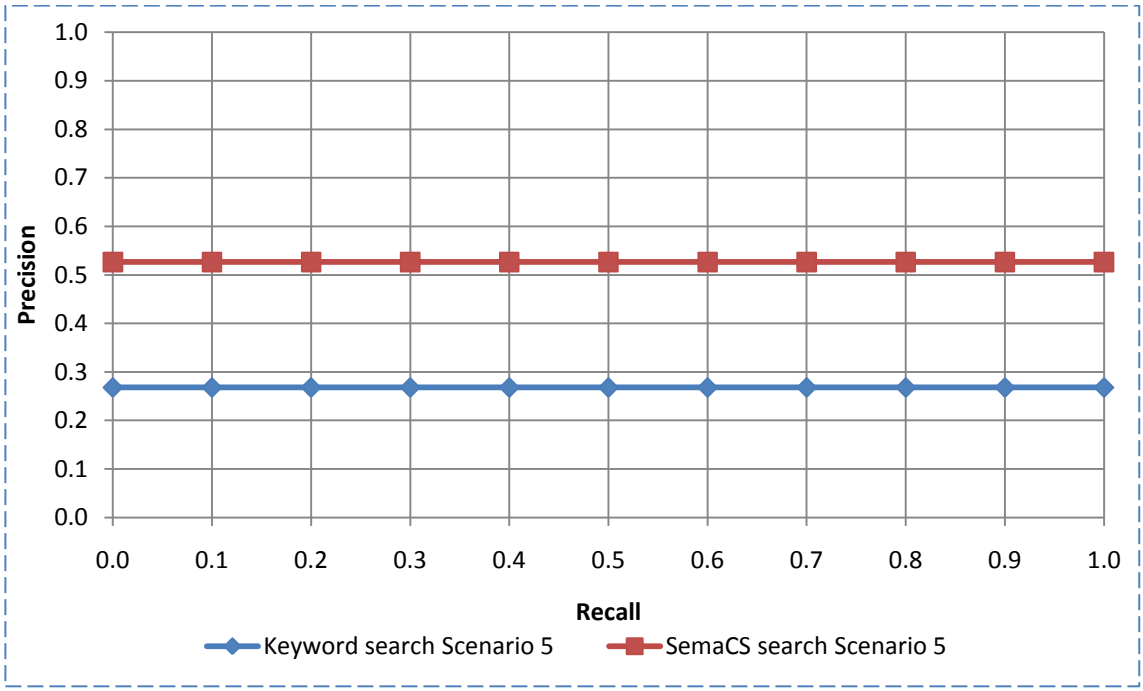

Figure 27: Scenario 5, 11-point interpolated average Precision Recall curves

- 5 scenario average

Through the 5 scenarios SemaCS has generally outperformed the Text-based implementation in overall Precision, Recall and F-score. However, SemaCS was consistently outperformed in MRR. This is made further evident via the combined experiment 11-point interpolated average Precision Recall curve (shown in Figure 28). It should be noted that 3 of the scenarios consisted of a

single correct answer; thus Recall could only take a value of 0% or 100%. Thus, the curves representing scenarios 2, 3 and 5 are linear. These linear curves also had a noticeable effect on the shape of the overall experiment curve shown in Figure 28.

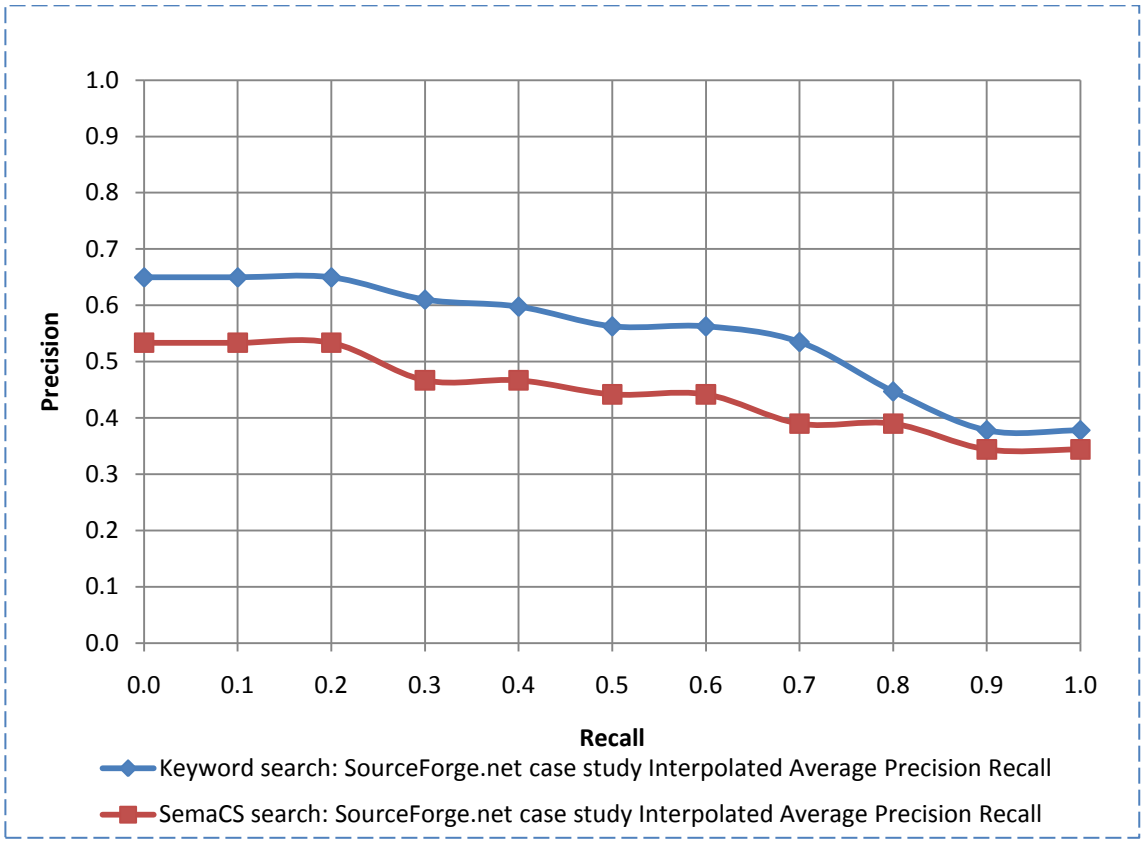

Figure 28: SourceForge.net study 11-point Interpolated average P/R curve

As made evident in Figure 28, SemaCS was clearly outperformed by the Keyword-based approach. This is the case because an 11-point interpolated Precision Recall curve is generated by only considering Precision at points of actual Recall. Such a curve also represents an ideal case where the result set ends with the last correctly matched element i.e. it does not take into account the total size of the result set. Yet a notable improvement of the SemaCS implementation over the Keyword-based implementation in actual Recall,

Precision and F-score is evident. These results indicate that SemaCS retrieval and matching algorithms performed as expected. However, the only conclusion that could account for both a decrease in performance evident in the 11-point interpolated average Precision Recall curves and an improvement in Recall, Precision and F-score (which are not oriented towards evaluation of relevance sorting approaches) is the fact that SemaCS result relevancy sorting algorithms did not perform as expected. This conclusion is further confirmed by the fact that SemaCS average MRR demonstrated a drop of 0.14 in ranking performance over the Keyword-based implementation. As a result of these findings, SemaCS result relevancy algorithms were improved before commencing the second study. Unlike with SourceForege.net study, with University of Westminster study (see next section) the actual strength of the match was used to make result relevancy decisions.

5.2.4 Summary

This section has presented and analysed results generated by the SourceForge.net case study. Although SemaCS MRR and 11-point interpolated average Precision Recall curves did not demonstrate an improvement on the Keyword-based approach, SemaCS was able to achieve a noticeable overall improvement of 11.61% in Precision, 6% in Recall and 12.49% in F-Score. By demonstrating these improvements, this study has provided positive supporting evidence for hypothesis $\mathbf{H_{1b}}$ (due to use of an Intranet based local data source to acquire semantic distance relations from) as well as further positive evidence supporting hypothesis $\mathbf{H_{1a}}$ (unlike pilot and second case studies, this study was implemented within software component domain). Additionally, further supporting evidence was provided for hypothesis $\mathbf{H_{1c}}$ because a domain taxonomy was automatically generated and successfully applied for query and domain interpretation.

This study has further identified that SemaCS prototype implementation result relevancy algorithms have not performed as expected. Although, the fact that a local dataset was used to extract mNGD (2) relation scores from is likely to have contributed towards a decrease of 0.14 in overall MRR score.

5.3 University of Westminster case study

University of Westminster SRS search system (see section 4.4 - Case study 2: University of Westminster SRS module search) was selected as a subject for this study because it provides a means to evaluate the primary and secondary hypothesis in an environment that is very different to the SourceForge.net study therefore providing further evidence to either support or negate primary hypothesis $\mathbf{H_1}$. Additionally, due to the different domain of application, this case study has also provided further positive evidence supporting hypothesis $\mathbf{H_{1a}}$. Furthermore, unlike SourceForge.net study, access to the WWW (via Yahoo! web interface) to acquire semantic distance relations was utilised hence providing further positive evidence supporting hypothesis $\mathbf{H_{1b}}$ in a different environment. Similarly to SourceForge.net study a single experiment was implemented (see section 4.5 - Case study 2: University of Westminster SRS module search) and a domain taxonomy was automatically generated providing further positive evidence supporting hypothesis $\mathbf{H_{1c}}$. Finally, this study was able to provide an easily managed data-based personalisation solution, hence providing evidence supporting hypothesis $\mathbf{H_{1d}}$. Two data sets have been used as part of the study – Year 2 modules (for Year 1 students to search – Appendix B1) and Year 3 modules (for Year 2 students to search - Appendix B2).

5.3.1 Year 1 students searching Year 2 modules

Year 1 study experiment was repeated 3 times using 3 implementations: University of Westminster SRS module search, SemaCS without data-based personalisation and SemaCS with data-based personalisation. Experiment procedures were designed in such a way as to allow for the removal of human factors in order to facilitate statistical evaluation and comparison via Precision, Recall, F-score and MRR (see section 4.5 'Case study 2: University of Westminster SRS module search'). However, it should be noted that SemaCS without data-based personalisation was subjected to the same restrictions as

SRS; only module titles were used to perform searches. This limitation was applied to allow for a meaningful comparison between the two approaches to take place. While SemaCS with data-based personalisation was not limited to module titles only; module descriptions were used to provide domain-wide data-based personalisation (used by SemaCS for search and domain taxonomy generation). In every other respect the three experiments were identical as they were performed using identical data sets and identical queries. A total of 22 subjects participated in the experiment (with participant queries collected and logged via the SemaCS implementation), 2 queries were found to be identical and one has been removed. The remaining 21 queries are shown in Table 11.

Q-ID	Student Query
1	oo programming
2	interface design
3	computing
4	graphics
5	web design
6	java
7	internet programming
8	java
9	c#
10	c# programming
11	Rapid Application Dev
12	Database systems
13	Graphics
14	internet
15	game
16	3d
17	.net
18	c++
19	INTERNET PROGRAMMING
20	.NET
21	mobile web xml xslt

Table 11: Case study 2 Year 1 experiment queries

The 21 participant queries shown in Table 11 were used to perform 21 searches on the 3 implementations: University of Westminster SRS module search, SemaCS without data-based personalisation and SemaCS with data-based personalisation (see Appendix B3 for corresponding experiment logs and B4 for calculation sheets). However, a concise version of the results is shown in Table 12. It should also be noted that, unlike with SourceForge.net case study, SemaCS did not implement a result cut-off strategy of any kind.

	SRS module search			SemaCS (no personalisation)			SemaCS with personalisation		
Q ID	Recall	Precision	RR	Recall	Precision	RR	Recall	Precision	RR
1	0.000000	0.000000	0.00	1.000000	0.833330	1.00	1.000000	0.714290	1.00
2	0.250000	1.000000	1.00	0.500000	0.250000	0.25	0.750000	0.130440	1.00
3	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
4	0.500000	1.000000	1.00	0.500000	1.000000	1.00	0.500000	1.000000	1.00
5	0.333330	1.000000	1.00	1.000000	0.111110	1.00	0.333330	0.062500	0.00
6	0.200000	1.000000	1.00	0.200000	0.333330	0.50	0.400000	1.000000	1.00
7	1.000000	1.000000	1.00	1.000000	0.200000	1.00	1.000000	0.142860	0.00
8	0.200000	1.000000	1.00	0.200000	0.333330	0.50	0.400000	1.000000	1.00
9	1.000000	1.000000	1.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
10	0.000000	0.000000	0.00	1.000000	0.200000	0.25	1.000000	0.153850	1.00
11	1.000000	1.000000	1.00	1.000000	0.166670	1.00	1.000000	0.076920	0.00
12	0.000000	0.000000	0.00	1.000000	0.200000	1.00	1.000000	0.142860	0.00
13	0.500000	1.000000	1.00	0.500000	1.000000	1.00	0.500000	1.000000	1.00
14	0.666670	1.000000	1.00	0.666670	1.000000	1.00	1.000000	1.000000	1.00
15	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
16	0.500000	1.000000	1.00	1.000000	1.000000	1.00	1.000000	1.000000	1.00
17	1.000000	1.000000	1.00	1.000000	1.000000	1.00	1.000000	1.000000	1.00
18	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.500000	0.026320	1.00
19	1.000000	1.000000	1.00	1.000000	0.200000	1.00	1.000000	0.142860	0.00
20	1.000000	1.000000	1.00	1.000000	1.000000	1.00	1.000000	1.000000	1.00
21	0.000000	0.000000	0.00	1.000000	0.096770	0.50	0.666670	0.333330	1.00

Table 12: Case study 2 Year 1 experiment results

The 21 automatically logged entries resulting from this experiment (given in Table 12) were analysed via an Excel spreadsheet (see Appendix B4) to arrive

at experiment F-score, MRR and average Precision and Recall values shown in Table 13.

	Precision	Recall	F-Score	MRR
SRS module search	66.67%	43.57%	52.70%	0.67
SemaCS (no personalisation)	42.50%	64.60%	51.27%	0.67
SemaCS (with personalisation)	47.27%	66.90%	55.40%	0.62

Table 13: Case study 2 Year 1 experiment results (average)

As shown in Table 13, SemaCS without data-based personalisation has demonstrated an average decrease of 24.17% in Precision and 1.43% in F-score when compared to the SRS module search. However, SemaCS without data-based personalisation has also achieved an average improvement of 21.03% in Recall while the MRR remained an identical 0.67. Although SemaCS with data-based personalisation has not managed to outperform the SRS implementation in Precision either, it has demonstrated a marginally smaller decrease of 19.4% while also displaying an increase of 23.33% in Recall and 2.7% in F-score. Using F-score as a base for comparison (see Figure 29) further demonstrates the differences between the 3 implementations. However, F-score is a set based measure of assessment and does not take into account the performance of the approach to score results. Nevertheless, it does provide a good estimate of the overall performance. As can be seen in Figure 29 SemaCS with data-based personalisation has achieved a noticeable improvement in overall performance. This provides positive supporting evidence for hypothesis $\mathbf{H_{1d}}$.

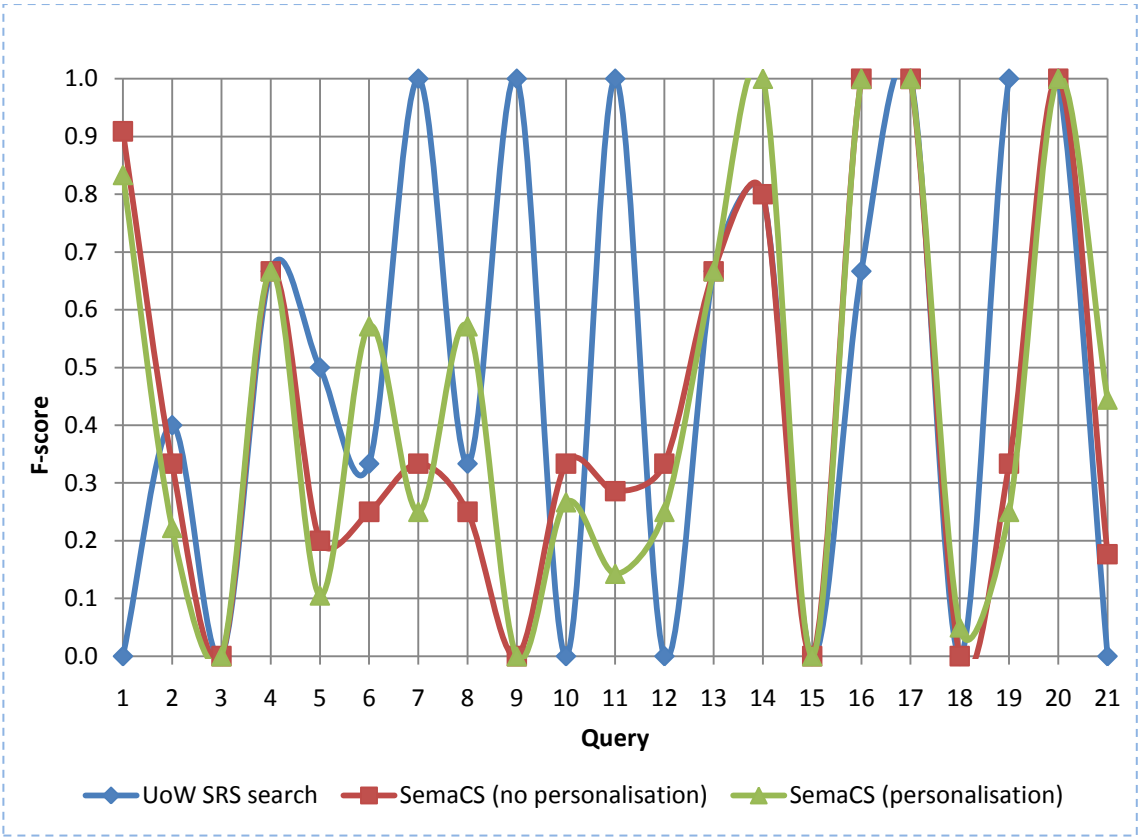

Figure 29: Case study 2 Year 1 per-query F-score comparison

It should also be noted that SemaCS with data-based personalisation has demonstrated a marginal decrease of 0.05 in MRR. This indicates that SemaCS was not as effective at ranking correct elements. However, an 11-point interpolated average Precision Recall comparison of the 3 implementations (see Figure 30) clearly demonstrates that SemaCS without data-based personalisation has performed better than the SRS. This is the case because with an 11-point interpolated average Precision Recall curve an ideal result set is considered (a cut-off point is the last returned relevant element) and given that SemaCS MRR is identical to SRS but Recall is higher – SemaCS has outperformed the SRS.

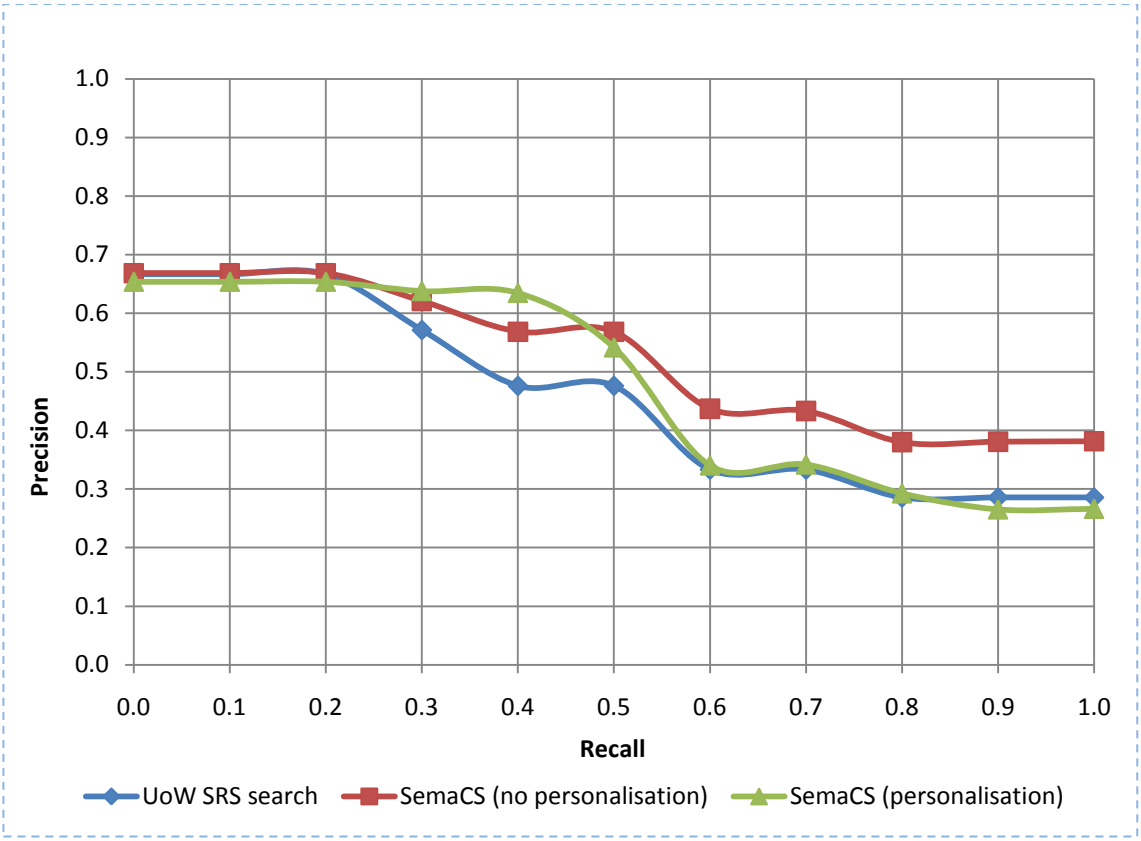

Figure 30: Case study 2 Year 1 Interpolated Average Precision Recall curve

However, SemaCS with data-based personalisation has not performed as expected. Although there is a clear improvement between recall points 0.3 and 0.4, its performance beyond these points is comparable to the SRS. Nonetheless, higher Recall levels were achieved when an entire result set is considered. Additionally, Precision scores were higher at recall points 0.3 and 0.4 (Figure 30) and thus evidence supporting hypothesis $\mathbf{H_{1d}}$ was provided.

Although neither SemaCS implementations have outperformed the SRS in Precision when the entire result set is considered, the fact that SemaCS has achieved an improvement in Recall and F-score provided further positive evidence in support of $\mathbf{H_1}$ and $\mathbf{H_{1d}}$ hypotheses and, in combination with previously collected data, positive evidence in support of $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$ and $\mathbf{H_{1c}}$.

Furthermore, it should be noted that SemaCS achieved these results without any expert input or training. It should also be noted that the SRS module search does appear to be extensively Precision biased. As demonstrated by the experiment logs (see Appendix B4F) SRS has either returned a correct result or no result at all. Thus, although achieving a very high Precision score, Recall levels were noticeably lower than any of the SemaCS implementations.

5.3.2 Year 2 students searching Year 3 modules

As with Year 1, Year 2 experiment was repeated 3 times using 3 different implementations: University of Westminster SRS module search, SemaCS without data-based personalisation and SemaCS with data-based personalisation. Experiment procedures were designed in such a way as to allow for the removal of human factors in order to facilitate statistical evaluation and comparison via Precision, Recall, F-score and MRR (see section 4.5 'Case study 2: University of Westminster SRS module search'). However, it should be noted that SemaCS without data-based personalisation was subjected to the same restrictions as SRS; only module titles were used to perform searches. This limitation was applied to allow for a meaningful comparison between the two approaches to take place. While SemaCS with data-based personalisation was not limited to module titles only; module descriptions were used to provide domain-wide data-based personalisation (used by SemaCS for search and domain taxonomy generation). In every other respect the 3 experiments were identical as they were performed using identical data sets and identical queries. A total of 29 subjects participated in this experiment (with participant queries collected via the SemaCS implementation), 10 queries were found to be identical and 1 was erroneous, thus 6 queries were removed. The remaining 23 queries are shown in Table 14.

Q-ID	Student Query
22	internet
23	ada
24	occam
25	business
26	Web Design
27	internet programming
28	graphics
29	web designing
30	website administration
31	learn flash
32	programming
33	networks
34	java games
35	html
36	java
37	flash design
38	network
39	graphic design
40	database design
41	3d design
42	database systems
43	computer graphics
44	graphics 3d multi

Table 14: Case study 2 Year 2 experiment queries

The 23 participant queries shown in Table 14 were used to perform 23 searches on the 3 implementations: University of Westminster SRS module search, SemaCS without data-based personalisation and SemaCS with data-based personalisation (see Appendix B3 for corresponding experiment logs and B5 for calculation sheets). However, a concise version of the results is shown in Table 15. It should also be noted that, unlike with SourceForge.net case study, SemaCS did not employ a result cut-off strategy (see section 4.2.1 - Precision) of any kind.

	SRS module search			SemaCS (no personalisation)			SemaCS with personalisation		
Q ID	Recall	Precision	RR	Recall	Precision	RR	Recall	Precision	RR
22	0.666670	1.000000	1.00	0.666670	1.000000	1.00	1.000000	1.000000	1.00
23	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
24	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
25	0.571430	1.000000	1.00	0.571430	1.000000	1.00	0.857140	1.000000	1.00
26	0.000000	0.000000	0.00	1.000000	0.073170	0.00	1.000000	0.111110	1.00
27	0.000000	0.000000	0.00	1.000000	0.333330	1.00	1.000000	0.200000	0.00
28	0.500000	1.000000	1.00	0.500000	1.000000	1.00	1.000000	1.000000	1.00
29	0.000000	0.000000	0.00	1.000000	0.071430	0.00	1.000000	0.500000	1.00
30	0.000000	0.000000	0.00	1.000000	0.500000	0.50	1.000000	0.058820	0.00
31	0.000000	0.000000	0.00	0.000000	0.000000	0.00	1.000000	0.250000	1.00
32	0.500000	1.000000	1.00	0.500000	0.750000	1.00	1.000000	0.857140	1.00
33	0.250000	1.000000	1.00	0.250000	1.000000	1.00	0.500000	0.800000	1.00
34	0.000000	0.000000	0.00	0.000000	0.000000	0.00	1.000000	0.285710	0.00
35	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
36	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.750000	1.000000	1.00
37	0.000000	0.000000	0.00	0.000000	0.000000	0.00	1.000000	0.035710	0.00
38	0.625000	1.000000	1.00	0.500000	1.000000	1.00	0.625000	1.000000	1.00
39	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
40	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
41	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.500000	0.027780	0.00
42	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.00
43	0.000000	0.000000	0.00	0.500000	0.333330	0.33	1.000000	0.285710	1.00
44	0.000000	0.000000	0.00	0.500000	0.052630	0.00	1.000000	0.133330	1.00

Table 15: Case study 2 Year 2 experiment results

The 23 automatically logged entries resulting from this experiment (given in Table 15) were analysed via an Excel spreadsheet (see Appendix B5) to arrive at experiment F-score, MRR and average values for Precision and Recall shown in Table 16. Unlike with Year 1 experiment, SemaCS without data-based personalisation has achieved an average increase of 4.84% in Precision when compared to the SRS module search. SemaCS without data-based personalisation has also achieved an average improvement of 21.19% in Recall and 14.9% in F-score as well as a 0.08 increase in MRR. SemaCS with data-based personalisation has outperformed the SRS system by a wider margin of

	Precision	Recall	F-Score	MRR
SRS module search	26.09%	13.54%	17.82%	0.26
SemaCS (no personalisation)	30.93%	34.73%	32.72%	0.34
SemaCS (with personalisation)	37.15%	66.23%	47.60%	0.52

Table 16: Case study 2 Year 2 experiment results (average)

11.06% in Precision, 52.69% in Recall, 29.78% in F-score and 0.26 in MRR. These improvements are made further evident via a per-query F-score comparison shown in Figure 31.

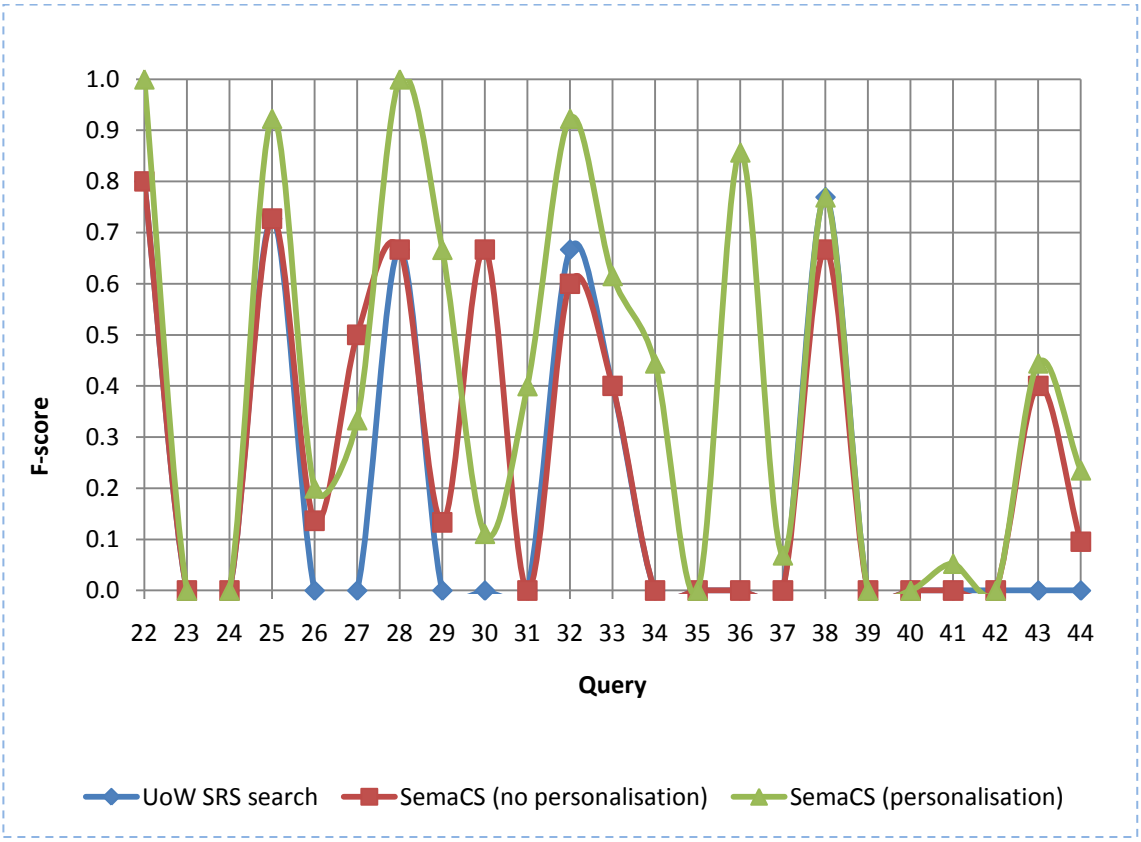

Figure 31: Case study 2 Year 2 per-query F-score comparison

However, it should be noted that although Year 2 queries (searching Year 3 modules) did not significantly differ form Year 1 queries (searching Year 2 modules), Year 3 modules are very subject specific, while year two modules, although more focused than year 1 modules, are still quite general. Consequently, it was a rarity for a participant query to be a direct keyword match to the module title. Thus the SRS, being keyword driven, has achieved a noticeably lower score compared to Year 1 experiment. These differences further made evident by the Year 2 experiment 11-point interpolated average Precision recall curve shown in Figure 32.

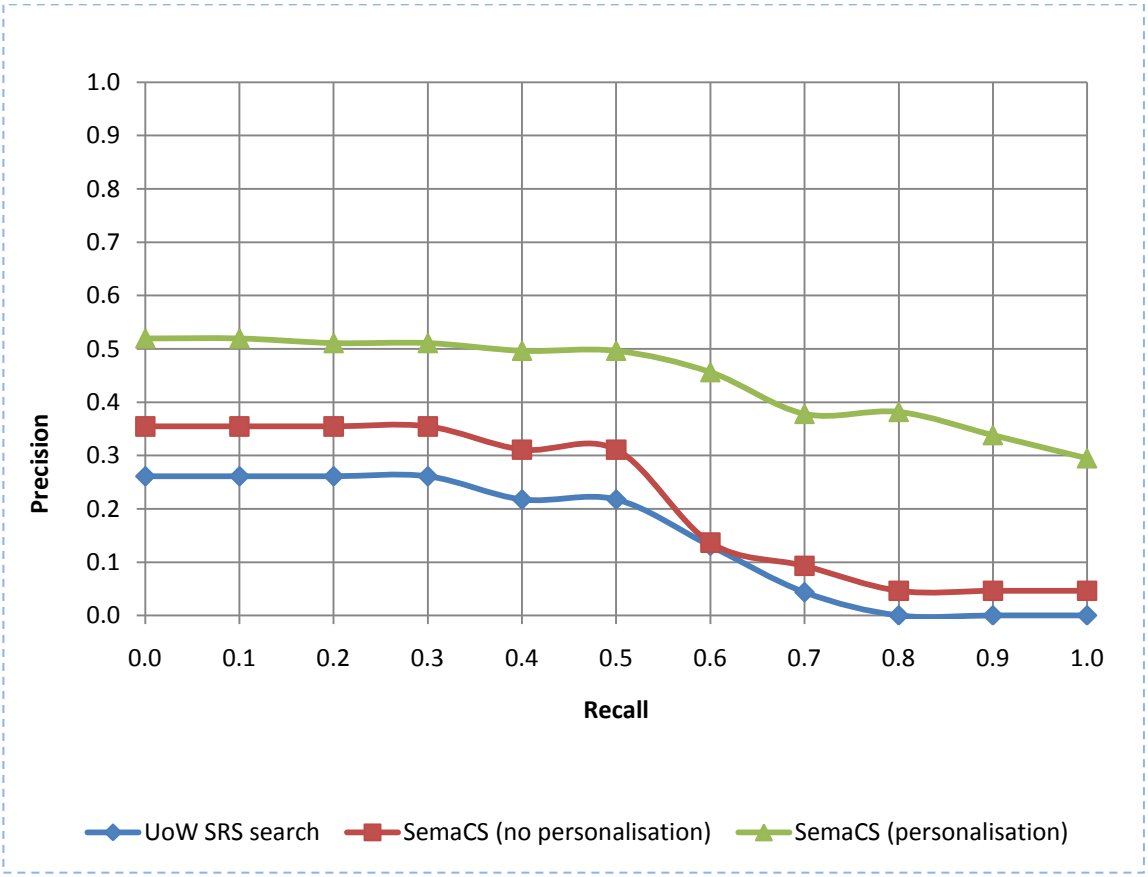

Figure 32: Case study 2 Year 2 interpolated average Precision Recall curves

As shown in Figure 32, both SemaCS implementations have managed to outperform the SRS by a noticeable margin. However, as was the case with Year 1 experiment, the SRS implementation has remained extremely Precision biased (see Appendix B5 F) and has either returned a correct match or no match at all. As both SemaCS implementations have outperformed the SRS in Precision, Recall, F-score and MRR this experiment has provided further positive evidence in support of $\mathbf{H_1}$ hypothesis and, in combination with previously collected data, positive evidence in support of $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$, $\mathbf{H_{1c}}$. Furthermore, as improvements in Precision, Recall, F-score and MRR were higher with data-based personalisation applied, further evidence supporting hypothesis $\mathbf{H_{1d}}$ was provided.

5.3.3 Combined Years 1 and 2 study result analyses

This study has provided positive evidence in support of $\mathbf{H_1}$ hypothesis. Web-sourced domain knowledge was automatically acquired and applied to aid with categorisation and search. Table 17 shows a combined average of Year 1 and Year2 experiments.

	Precision	Recall	F-Score	MRR
SRS module search	46.38%	28.55%	35.35%	0.46
SemaCS (no personalisation)	36.71%	49.67%	42.22%	0.50
SemaCS (with personalisation)	42.21%	66.57%	51.66%	0.57

Table 17: Case study 2 Years 1 and 2 experiment results (average)

Contrary to SourceForge.net study experiment where high Recall levels were achieved by the Text-based implementation, SRS (being likewise Text-based) has only achieved an average Recall of 28.55%. Such a drop in performance can be partially explained by the fact that more ‘realistic’ queries were provided

by participants. Furthermore, these real queries were not directed by the experts in any way – participants searched for modules without being provided with scenarios or any specific task. Additionally, the SRS system only uses module titles to carry out its search (module descriptions are not searched). And, because Year 3 modules are very subject specific, it was a rarity for a participant query to be a direct keyword match to the module title. Unlike the SRS SemaCS has an ability to perform indirect matches and, as a direct result of this ability, has achieved an improvement in Recall and F-score thus providing positive evidence in support of $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$ and $\mathbf{H_{1c}}$ hypotheses as well as data-based personalisation – $\mathbf{H_{1d}}$. However, as a 2 Year average, SemaCS was not able to outperform the SRS in Precision score. Nonetheless, it is feasible that higher Precision levels could be achieved by sacrificing a measure of Recall. Furthermore, as demonstrated by the last experiment, Precision and Recall levels can also be improved by providing greater levels of data-based personalisation. It is likewise feasible that Precision levels similar to that of the SRS system would have been achievable by simply ensuring that all modules have a description (currently about 20% do not, although this is the limitation introduced by the SRS) or implementing a cut-off rule. Figure 33 represents an effect of such a cut-off rule, an 11-point interpolated average Precision Recall curve only conceders an ideal result set where the last element of the result set is the last matching element. When such an ideal result set is considered, both SemaCS implementations clearly outperform the SRS. Furthermore, SemaCS with data-based personalisation outperforms both the SemaCS and the SRS implementations by a clear margin.

5.3.4 Summary

This section has presented and analysed results generated by the University of Westminster case study. These results demonstrate that SemaCS without data-based personalisation achieved a Precision score 9.67% less than that achieved by SRS. However, an improvement of 21.11% in Recall score and 6.87% in F-

score was demonstrated. This was expected as only module titles have been used to perform searches (this is a limitation of the SRS search system).

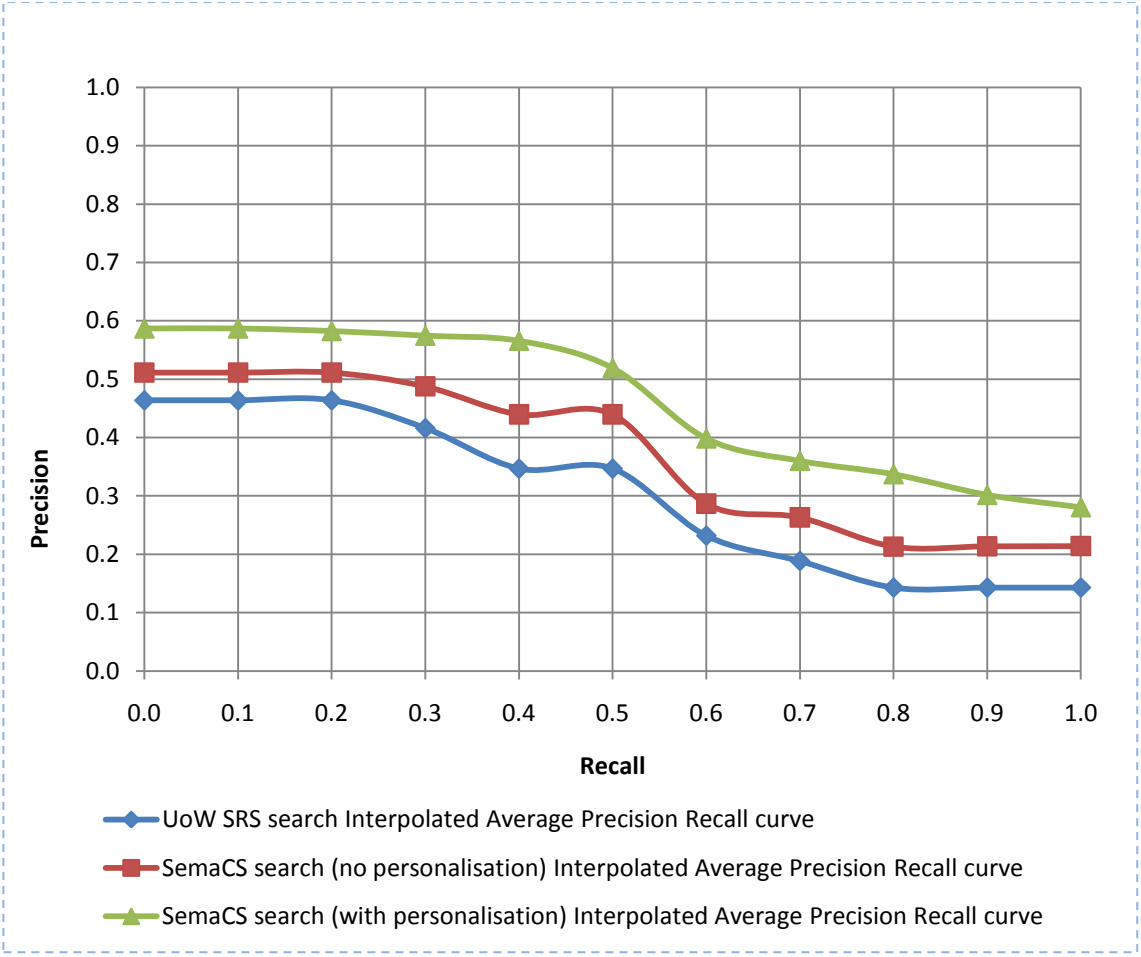

Figure 33: Case study 2 Years 1 and 2 interpolated average P/R curves

Consequently, with data-based personalisation applied, SemaCS has demonstrated an improvement of 38.01% in Recall and 16.32% in F-score. Although, even with data-based personalisation, an improvement in Precision could not be achieved. Nevertheless, by demonstrating these improvements this study has provided further positive supporting evidence for primary and secondary hypothesis and specifically for hypothesis $\mathbf{H_{1d}}$ as data-based personalisation has achieved a significant improvement in result relevancy.

Finally, it should be noted that although experiment 11-point interpolated average Precision Recall curves provide a means of comparing the 3 implementations, their shape is unusual. This is the case because such curves are best suited to evaluate an approach over large document collections and many queries that have multiple answers. Furthermore, as neither SemaCS nor SRS display the whole result set, it was common that not all relevant elements were returned. Due to these reasons, experiment curves did not start at Precision position 1 nor did they gracefully diminish towards Recall position 1.

5.4 Conclusion

The pilot case study assessed the primary and secondary hypotheses. This study has provided partial evidence in support of $\mathbf{H_1}$ hypothesis by providing positive evidence related to mNGD (2). Additionally, partial evidence supporting $\mathbf{H_{1a}}$ hypothesis was provided as the terms used in the data set were not related and belonged to a multitude of domains and concepts.

The pilot study was shortly followed by SourceForge.net and University of Westminster SRS case studies. SourceForge.net study was orientated towards qualitative methodology (due to a small number of scenarios) and University of Westminster towards a quantitative methodology (due to unlimited number of possible scenarios). As both studies are distinct and based within different application domains both were better suited to evaluate $\mathbf{H_1}$ and $\mathbf{H_{1a}}$ hypothesis than either on their own. Although study subjects were distinct, evaluation criteria have remained uniform throughout. Study results were analysed to produce an average figure for Precision and Recall as well as experiment MRR and F-score. Results generated by both studies demonstrate a clear improvement in results when compared to Text-based search and SRS implementations and therefore demonstrate that application of automatically sourced web knowledge has a positive impact on results. We can therefore

conclude that, although results were not scored as high as expected, evidence validating hypotheses **H_1**, **H_{1a}**, **H_{1b}**, **H_{1c}**, and **H_{1d}** were provided.

Chapter 6 Conclusions and future work

The aim of this thesis was to define a framework capable of providing a means of automated software component categorisation and selection utilising domain knowledge obtained from textual descriptions of software components being categorised and the WWW. This objective was realised by defining SemaCS – an automated categorisation and selection framework that does not rely on expert-driven annotation or content generation.

Although originally SemaCS was to provide this solution for the software component domain, the means of processing and understanding natural text within the software component domain are applicable to many other domains. Consequently, this research focused on domain independent automated means of categorisation and search of natural text descriptions. This objective was phrased as hypothesis $\mathbf{H_1}$.

However, $\mathbf{H_1}$ hypothesis represents a complex problem which is difficult to evaluate. To further aid with validation, $\mathbf{H_1}$ hypothesis was split into a set of smaller secondary hypothesis represented as:

- Domain independent (based purely on free-form textual descriptions and access to the WWW) secondary hypothesis $\mathbf{H_{1a}}$:

- Semantically driven (by means of scale invariant modified NGD to acquire semantic distances) secondary hypothesis $\mathbf{H_{1b}}$:

- Capable of automated domain taxonomy generation and search (as a result removing the need for expert training or manual annotation) secondary hypothesis $\mathbf{H_{1c}}$:

- Allowing for data-based personalisation (consequently improving result relevancy for given task/domain) secondary hypothesis $\mathbf{H_{1d}}$:

6.1 mNGD modification

SemaCS utilises domain knowledge obtained from the objects being categorised and the WWW (or any other sources such as company documentation). SemaCS further implements a modified version of the NGD (Cilibrasi and Vitanyi 2007) algorithm, mNGD (2), to detect a degree of relatedness between words, which is used to make relevancy decisions. The modification comes in the form of **N** originally representing total number of pages referenced by Google (Cilibrasi and Vitanyi 2004). For SemaCS purposes dependability on unpredictable scale variable **N** value is removed. mNGD (2) was evaluated within three studies. And, based on statistical study data analyses, it can be concluded that mNGD (2), unlike NGD (1), is able to acquire semantic distances without depending on N – hypothesis $\mathbf{H_{1b}}$ has been validated.

mNGD (2) modification provides an ability to acquire semantic distances without depending on a changing and unpredictable external factor: the number of documents in the collection. This modification further allows for dynamically changing document collections to be used (for example data-based personalisation) without influencing the process or significance of detection. Not having to rely on this value also assures that relation weights are calculated for the sub-set in which both terms are discovered (the semantic relation discovered between terms **x** and **y** does not depend on how many documents there are in

the collection). This is useful when working with rare terms or small document collections (or collections where size is not known or changes rapidly).

6.2 Evaluation criteria discussion

It should be noted that the chosen study evaluation criteria are not ideally suited to evaluate SemaCS. SemaCS results are scored based on a relation that is not always of a type that the originator of the query assumes. Some kind of a relation to the data and query exists, even if no hard match is detected based on expert judgement. This is both a benefit and a failure: as a benefit SemaCS is able to forgo syntax mistakes and still detect a relation that is almost as strong as having used the correct spelling; as a drawback such relations (or even those that seem invalid in the originators view) are also detected. With semantic approaches, as with human logic in general, everything is not either true or false; there are intermediate concepts (e.g., maybe, a little, quite a lot, etc.). Consequently, results returned by SemaCS cannot be analysed accurately by applying simple Boolean evaluation. This shortfall has been recognised. However, the second case study has offered a possible solution. Although the same evaluation criteria were used, a significant difference was introduced in the way scenarios were created and assigned: with SourceForge.net study there was a fixed predefined number of expert generated scenarios; this was not the case with SRS study where participants were allowed to ‘free search’ for anything of interest within the study domain. These differences provide for a more realistic representation of the real world requirements and, as a result, a more realistic means of evaluation.

6.3 Main findings of the thesis

One feasibility and two primary case studies were implemented to validate the ideas and hypothesis proposed in this thesis: Multi-domain pilot study,

SourceForge.net study, and University of Westminster SRS search study.

Pilot study has provided positive evidence supporting hypothesis $\mathbf{H_{1b}}$, mNGD (2), and consequently providing positive evidence supporting primary $\mathbf{H_1}$ hypothesis. This study has also identified a close association between mNGD (2) and NGD (1) generated scores. However, due to its low (or nonexistent) impact on $\mathbf{H_1}$ hypothesis and resource limitations this association was not investigated further.

SourceForge.net study was orientated towards qualitative methodology (due to a small number of scenarios) and University of Westminster towards a quantitative methodology (due to unlimited number of possible scenarios). As both studies are distinct and based within different application domains both were better suited to evaluate primary and secondary hypothesis than either on their own.

Based on statistical analyses of these study result analyses, it can be concluded that secondary hypothesis $\mathbf{H_{1a}}$, $\mathbf{H_{1b}}$, $\mathbf{H_{1c}}$ and $\mathbf{H_{1d}}$ are validated, therefore, it can also be concluded that hypothesis $\mathbf{H_1}$ is validated. However, it can also be concluded that SemaCS, in its current form, is not able to compete with manually trained systems at the same level of accuracy. This is the case because SemaCS only achieved ≈ 67% Recall. Nevertheless, this level of Recall was achieved without any expert input or expert defined structure. Additionally, SemaCS dealt with complex, erroneous queries and natural language descriptions without receiving any training. Furthermore, SemaCS has exceeded a Keyword-based SRS implementation by a considerable amount when query interpretation was required (see section 5.3.2 Year 2 students searching Year 3 modules).Nonetheless, although SemaCS has demonstrated higher levels of

Recall as well as an ability to interpret user queries, the MRR were generally lower than that of the Text-based implementations. Thus it can also be concluded that SemaCS, in its current form, is best applied to search, user query interpretation and domain description tasks.

6.4 Implications for the field

This research is significant for several reasons. Firstly, it proposed a novel approach to the implementation of component repositories and retrieval of appropriate reusable software components. Secondly, a novel approach to semantic domain taxonomy generation was defined. Lastly, the SemaCS approach to information discovery utilises cross-domain architecture that can be deployed without any training.

Consequently, this research benefits general users as well as application designers because it defines a novel software component categorisation approach that can provide better matched results then current traditional textual or statistically bound approaches. Furthermore, as a novel approach to extraction of semantic information is proposed, there exists an impact on IR domain and, specifically, on Clustering, word disambiguation, NLP and Web/Intranet search engines.

6.5 Directions for future work

One of the issues identified with SemaCS implementation was the amount of time it required to generate domain taxonomy and textual description indices. Thus, currently, SemaCS cannot be scaled to large datasets (like the WWW). However, this problem could be addressed by forgoing the need to keep document collection indices, as was originally intended. In this case, search queries would be matched against the domain taxonomy to generate a search

taxpet (which can also be enriched with related terms/words defined in the taxonomy) and then intersected (or matched using basic keyword matching algorithms) with textual descriptions. As a result, efficiency of keyword matching algorithms can be achieved at the search stage with a small overhead introduced at the query expansion/processing (especially if data-based personalisation is used to enrich the query) and domain taxonomy generation stages. Although a model of the domain is still generated, it is (compared to creating an index of all terms within all descriptions) a very small overhead.

A further improvement to SemaCS model generation efficiency can be achieved by using further functionality provided by GATE (Cunningham, Maynard et al. 2002). In addition to common POS (only functionality currently used) GATE is capable of regressing (stemming) terms being processed to a root form (for example, did, do, done etc. should be regressed to the same term ‘do’).

Further evaluation of SemaCS can also be performed. A popular approach is to use universal data collections (for example Text Retrieval Conference (TREC) Web or Terabyte collections), because these collections generally contain documents as well as queries and answers. Although, similarly to SemaCS case studies, queues and answers are person generated, because collections are used in a number of related evaluations, a detailed comparison to other approaches could be made.

Further research related to SemaCS result element relevancy decision algorithms can be performed. Although current algorithms are well suited to retrieval, query interpretation and domain description tasks, they are not as efficient at deciding result element relevancy. It is possible that, due to the inherent limitations introduced by mNGD (2), SemaCS algorithms cannot be

improved. However, user level personalisation, as well as non SemaCS-based algorithms could be applied for such a purpose.

As SemaCS was designed to create a domain taxonomy by detecting in-use word patterns (i.e. words that are commonly used in data collection form Tier1 elements etc.). SemaCS could be employed as a means of automated detection of repeating patterns such as service descriptions (e.g. WSDL (W3C 2001) or WSMO (Roman, Keller et al. 2005)).

SemaCS could also be adapted to provide for context-based navigation. The concept itself is interpreted differently and can imply personalisation or document based focusing and interpretation of the query as well as non textual query browsing of concepts and hierarchies or just topic specific search (see (Finkelstein, Gabrilovich et al. 2001) for a more detailed introduction). However, the type of taxonomy generated by SemaCS algorithms cannot be directly applied for user navigation purposes. Nevertheless, because the basic idea behind context-based navigation could be easily captured by SemaCS data-based personalisation and algorithms, a further direction could be adaptation of the SemaCS framework to provide for such functionality. This could be achieved through reuse of existing expert generated taxonomies or, perhaps, as a hybrid approach only providing mNGD (2) based context aware matching and interpretation.

Finally, mNGD (2) algorithm could be applied to other ML, DM and NLP approaches. Additionally, SemaCS framework can be implemented using other approaches of domain taxonomy generation.

6.6 Conclusion

Having performed the three studies and analysed their results, it can be concluded the $\mathbf{H_1}$ hypothesis is validated and useful information aiding in search and categorisation can be automatically extracted from within the public domain or even a localised corpus such as project documentation or Intranet pages. As a result it can also be concluded that SemaCS is able to achieve these results with a smaller footprint than manually trained approaches. Furthermore, as SemaCS could be adapted to function within large datasets (for example, with WWW) it also presents an improvement on automated approaches.

Nevertheless, it can also be concluded that SemaCS, in its current form, is not able to compete with manually trained systems at the same level of accuracy. This is the case because SemaCS, at best, was able to achieve ≈ 67% Recall while some manually trained (and indeed automated) approaches are able to achieve 80% or more. Yet, SemaCS has received no training, nor was the prototype fully implemented. Additionally, SemaCS dealt with complex, erroneous queries and natural language descriptions.

SemaCS was able to perform its function without relying on explicit expert input. In this dimension it exceeds any manually configured and maintained approach. It can also be concluded that SemaCS is able to function better with personalisation. Furthermore, higher levels of flexibility and results relevancy could be achieved by implementing user-level (profile) data-based personalisation. We cannot say that SemaCS is perfectly suited to the domains in question but it does provide proof of concept and, given further improvements are investigated and incorporated, has the potential to approach accuracy levels of manually trained systems while remaining efficient and scalable to large document collections.

References and bibliography

Aberer, K., T. Catarci, et al. (2004). Emergent Semantics Systems. 1st International IFIP Conference, Paris France, Springer-Verlag.

Abts, C. (2002). COTS-Based Systems (CBS) Functional Density A Heuristic or Better CBS Design. 1st International Conference on COTS-Based Software Systems, Orlando FL USA, Springer-Verlag.

Alani, H., S. Kim, et al. (2003). "Automatic Ontology-Based Knowledge Extraction from Web Documents." IEEE Intelligent Systems 18(1): 14-21.

Albert, C. and L. Brownsword (2002). Meeting the Challenges of Commercial-Off-The-Shelf (COTS) Products: The information Technology Solutions Evolution Process (ITSEP). 1st International Conference on COTS-Based Software Systems, Orlando FL USA, Springer-Verlag.

Alves, C. and J. Castro (2001). CRE: A Systematic Method for COTS Component Selection. 15th Brazilian Symposium on Software Engineering, Rio de Janeiro Brazil.

Alves, C. and A. Filkelstein (2002). Challenges in COTS Decision-Making: A Goal-Driven Requirements Engineering Perspective. 14th International Conference on Software Engineering and Knowledge Engineering, Ischia Italy, ACM Press.

Anyanwu, K., A. Maduko, et al. (2005). SemRank: Ranking Complex Relationship Search Results on the Semantic Web. 14th international conference on World Wide Web, Chiba Japan, ACM Press.

Apache. (2006). "Apache Lucene open source project." Retrieved 5 Feb 2006, from http://lucene.apache.org/java/docs/.

Atkinson, S. (1998). Modelling formal integrated component retrieval. 5th International Conference on Software Reuse, Victoria B.C. Canada, IEEE Computer Society Press.

Ayala, C. P., P. Botella, et al. (2005). On Goal-Oriented COTS Taxonomies Construction. 4th International Conference on COTS-Based Software Systems, Bilbao Spain, Springer-Verlag.

Baeza-Yates, R. A. and B. A. Ribeiro-Neto (1999). Modern Information Retrieval, ACM Press/Addison-Wesley.

Baker, T. G. (2002). Lessons Learned Integrating COTS into Systems. 1st International Conference on COTS-Based Software Systems, Orlando FL USA, Springer-Verlag.

Ballurio, K., B. Scalzo, et al. (2002). Risk Reduction in COTS Software Selection with BASIS. 1st International Conference on COTS-Based Software Systems, Orlando FL USA, Springer-Verlag.

Bandini, S., F. D. Paoli, et al. (2002). A support system to COTS-based software development for business services. 14th International Conference on Software Engineering and Knowledge Engineering, Ischia Italy, ACM Press.

Bar-Ilan, J. (1998). "Search engine results over time: a case study on search engine stability." International Journal of Scientometrics, Informetrics and Bibliometrics 2(1): 1-16.

Barbier, F. (2004). Web-Based COTS Component Evaluation. 3rd International Conference on COTS-Based Software Systems, Redondo Beach CA USA, Springer-Verlag.

Benbasat, I., D. K. Goldstein, et al. (1987). "The Case Research Strategy in Studies of Information Systems." MIS Quarterly 11(3): 369-386.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques Grouping Multidimensional Data Recent Advances in Clustering J. Kogan, C. Nicholas and M. Teboulle, Springer-Verlag**:** 25-71.

Berners-Lee, T. (1989). "Information Management: A Proposal." Retrieved 2/04/2008, from http://www.w3.org/History/1989/proposal.html.

Berners-Lee, T. and M. Fischetti (1999). Weaving the Web, Harper San Francisco.

Bettstetter, C. and C. Renner (2000). A Comparison of service discovery protocols and implementation of the service location protocol. 6th EUNICE Open European Summer School, Twente Netherlands.

Beus-Dukic, L. and J. Bøegh (2003). COTS Software Quality Evaluation. 2nd International Conference on COTS-Based Software Systems, Ottawa Canada, Springer-Verlag.

Bontcheva, K., V. Tablan, et al. (2004). "Evolving GATE to Meet New Challenges in Language Engineering." Natural Language Engineering 10: 349-373.

Brin, S. and L. Page (1998). "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30(1-7): 107-117.

Broder, A. (2002). "A taxonomy of web search." ACM SIGIR Forum 36(2): 3-10.

Brown, A. B. (2000). Large Scale Component-Based Development Prentice Hall.

Brown, A. W. and K. Short (1997). On components and objects: the foundations of component-based development. 5th International Symposium on Assessment of Software Tools, Pittsburgh PA USA, IEEE Computer Society Press.

Brown, A. W. and K. C. Wallnau (1996). Engineering of component-based systems. 2nd IEEE International Conference on Engineering of Complex Computer Systems, Montreal Que. Canada, IEEE Computer Society Press.

Brownsword, L. L., D. J. Carney, et al. (2004). Current Perspectives on Interoperability, Software Engineering Institute, Carnegie Mellon University.

Buckley, C., G. Salton, et al. (1994). The effect of adding relevance information in a relevance feedback environment. 17th Annual International SIGIR Conference on Research and Development in Information Retrieval, Dublin Ireland, ACM Press.

Budanitsky, A. and G. Hirst (2006). "Evaluating WordNet-based Measures of Lexical Semantic Relatedness." Computational Linguistics 32(1): 13-47.

Buntine, W., J. Lofstrom, et al. (2004). A Scalable Topic-Based Open Source Search Engine. IEEE/WIC/ACM International Conference on Web Intelligence, Beijing China, IEEE Computer Society Press.

Capra, R. G., III and M. A. Perez-Quinones (2005). "Using Web search engines to find and refind information." Computer 38(10): 36-42.

Carney, D. and F. Leng (2000). "What do you mean by COTS? Finally, a useful answer." IEEE Software 17(2): 83-86.

Carney, D. J., E. J. Morris, et al. (2003). Identifying Commercial Off The Shelf (COTS) Product Risks: The COTS Usage Risk Evaluation, Software Engineering Institute, Carnegie Mellon University.

Carr, L., W. Hall, et al. (2001). Conceptual Linking: Ontology-based Open Hypermedia. 10th international conference on World Wide Web, Hong Kong, ACM Press.

Carvallo, J. P., X. Franch, et al. (2004). On the Use of Quality Models for COTS Evaluation. 26th International Conference on Software Engineering, Edinburgh Scotland, The IET.

Carvallo, J. P., X. Franch, et al. (2003). Defining a Quality Model for Mail Servers. 2nd International Conference on COTS-Based Software Systems, Ottawa Canada, Springer-Verlag.

Cechich, A. and M. Piattini (2004). On the Measurement of COTS Functional Suitability. 3rd International Conference on COTS-Based Software Systems, Redondo Beach CA USA, Springer-Verlag.

Cechich, A. and M. Piattini (2005). Filtering COTS Components Through an Improvement-Based Process. 4th International Conference on COTS-Based Software Systems, Bilbao Spain, Springer-Verlag.

Ceravolo, P. and E. Damiani (2003). A Ontology-based Process Modelling for XP. 10th Asia-Pacific Software Engineering Conference, Chiangmai Thailand, IEEE Computer Society Press.

Chen, H. and S. Dumais (2000). Bringing order to the Web: Automatically categorizing search results. SIGCHI Conference on Human Factors in Computing Systems, New York NY, ACM Press.

Chiang, C.-C. (2002). Development of Reusable Components through the Use of Adapters. 36th Annual Hawaii International Conference on System Sciences, Hilton Waikoloa Village Island of Hawaii, IEEE Computer Society Press.

Cilibrasi, R. L. and P. M. B. Vitanyi. (2004). "Automatic Meaning Discovery Using Google." Retrieved 26 Dec. 2006, from http://xxx.lanl.gov/abs/cs.CL/0412098.

Cilibrasi, R. L. and P. M. B. Vitanyi (2007). "The Google Similarity Distance." IEEE Transactions on Knowledge and Data Engineering 19(3): 370-383.

CLARIFI. (2007). "CLARIFI project." Retrieved 14 Dec. 2008, from http://clarifi.eng.it/.

Clark, B. and M. Torchiano (2004). COTS Terminology and Categories: Can We Reach a Consensus? 3rd International Conference on COTS-Based Software Systems, Redondo Beach, CA, USA, Springer-Verlag.

Clements, P. C. (1995). "From Subroutines to Subsystems: Component-Based Software Development." American Programmer 8(11): 31-39.

Clusty. (2004). "Clusty - the clustering search engine." Retrieved 21 Jun. 2007, from http://clusty.com/.

ComponentSource. (1996). "ComponentSource Software component search engine homepage." Retrieved 13 Jan 2008, from http://www.componentsource.com.

Consroe, K. (1999). "In Memoriam." Retrieved 1st Feb 2009, from http://www.cs.cornell.edu/Info/Department/Annual96/Beginning/salton.html.

Crnkovic, I., J. Axelsson, et al. (2005). COTS Component-Based Embeded Systems - A Dream or Reality. 4th International Conference on COTS-Based Software Systems, Bilbao Spain, Springer-Verlag.

Crnkovic, I., S. Larson, et al. (2002). "Component-based software engineering: building systems from components." ACM SIGSOFT Software Engineering Notes 27(3): 47-50.

Crnkovic, I. and M. Larsson (2004). Classification of quality attributes for predictability in component-based systems. DSN Workshop on Architecting Dependable Systems, Florence Italy.

Crnkovic, I., H. Schmidt, et al. (2004). "6th ICSE Workshop on Component-Based Software Engineering: Automated Reasoning and Prediction." ACM SIGSOFT Software Engineering Notes 29(3): 1-7.

Cunningham, H., D. Maynard, et al. (2002). "The GATE User Guide." Retrieved 14 Mar. 2007, from http://gate.ac.uk/sale/tao/split.html.

Cunningham, H., D. Maynard, et al. (2002). "GATE - General Architecture for Text Engineering homepage." Retrieved 16 Sep 2007, from http://gate.ac.uk/.

Cunningham, H., D. Maynard, et al. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. 40th Annual Meeting on

Association for Computational Linguistics, Philadelphia Pennsylvania, Association for Computational Linguistics.

Cunningham, H., D. Maynard, et al. (2002). GATE: an Architecture for Development of Robust HLT Applications. 40th Annual Meeting on Association for Computational Linguistics, Philadelphia Pennsylvania, Association for Computational Linguistics.

Cunningham, H., D. Maynard, et al. (2006). "Developing Language Processing Components with GATE Version 3 (a User Guide)." Retrieved 17 Sep 2007, from http://gate.ac.uk/sale/tao/index.html.

Cutting, D. R., D. R. Karger, et al. (1992). Scatter/Gather: a cluster-based approach to browsing large document collections. 15th annual international ACM SIGIR conference on Research and development in information retrieval, Copenhagen Denmark, ACM Press.

Daoud, M., L. Tamine-Lechani, et al. (2009). A session based personalized search using an ontological user profile. ACM symposium on Applied Computing, Honolulu Hawaii, ACM Press.

Dashofy, E. M., A. v. d. Hoek, et al. (2002). An infrastructure for the rapid development of XML-based architecture description languages. 24th International Conference on Software Engineering, Orlando Florida, ACM Press.

DCMI. (2007). "The Dublin Core Metadata Initiative." Retrieved 17 May 2007, from http://dublincore.org/.

Deerwester, S., S. T. Dumais, et al. (1990). "Indexing By Latent Semantic Analysis." Journal of the American Society for Information Science and Technology 41: 391-407.

Delic, K. A. and B. Hoellmer (2000). "Knowledge-based support in help-desk environments." IT Professional 2(1): 44-48.

Ding, L., T. Finin, et al. (2004). Swoogle: a search and metadata engine for the semantic web. 13th ACM International Conference on Information and Knowledge Management Washington D.C. USA ACM Press.

Ding, L., T. Finin, et al. (2005). "Search on the Semantic Web." Computer 38(10): 62-69.

Doerr, M. (2003). "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata." AI Magazine 24(3): 75-92.

Dowman, M., V. Tablan, et al. (2005). Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News. 14th International World Wide Web Conference, Chiba Japan, ACM Press.

Dridi, O. (2008). Ontology-based information retrieval: Overview and new proposition. 2nd International Conference on Research Challenges in Information Science, Marrakech Morocco, IEEE Computer Society Press.

DSE. (2005). "DSE - Dedicated Systems Encyclopaedia." Retrieved 19 Nov 2006, from http://www.omimo.be/encyc/.

eCots. (2005). "COTS software component search platform." Retrieved 10 Dec. 2006, from http://www.ecots.org.

Edgington, T., T. S. Raghu, et al. (2005). Knowledge Ontology: A Method for Empirical Identification of 'As-Is' Contextual Knowledge. 38th Hawaii International Conference on System Sciences, Hilton Waikoloa Village Island of Hawaii, IEEE Computer Society Press.

Egashira, R., A. Enomoto, et al. (2004). Distributed and Adaptive Discovery Using Preference. International Symposium on Applications and the Internet, Tokyo Japan, IEEE Computer Society Press.

Elgazzar, S., A. Kark, et al. (2005). COTS Acquisistion: Getting a Good Contract. 4th International Conference on COTS-Based Software Systems, Bilbao Spain, Springer-Verlag.

Emtage, A. and P. Deutsch (1992). Archie - An Electronic Directory Service for the Internet. USENIX Winter 1992 Technical Conference, San Francisco USA, USENIX Association.

Erdmann, M., A. Maedche, et al. (2000). "From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools." from http://www.uni-koblenz.de/~staab/Research/Publications/erdmannetal_semann2000.pdf.

Erofeev, S. and P. D. Giacomo (2006). Usage of Dynamic Decision Models as an Agile Approach to COTS Taxonomies Construction. 5th International Conference on COTS-Based Software Systems Orlando Florida, IEEE Computer Society Press.

Ester, M., H.-P. Kriegel, et al. (1998). Incremental Clustering for Mining in a Data Warehousing Environment. 24th International Conference on Very Large Data Bases, New York USA, Morgan Kaufmann Publishers Inc.

Ferragina, P. and A. Gulli (2005). A personalized search engine based on Web-snippet hierarchical clustering. 14th International World Wide Web Conference, Chiba Japan, ACM Press.

Finkelstein, L., E. Gabrilovich, et al. (2001). Placing search in context: the concept revisited. 10th international conference on World Wide Web Hong Kong.

Flamenco. (2007). "The Flamenco Search Interface Project." Retrieved 3 Aug 2007, from http://flamenco.berkeley.edu/pubs.html.

Flashline. (2005). "Flashline - Software component search engine homepage." Retrieved 9 Sep. 2006, from http://www.flashline.com.

Frank, E., M. Hall, et al. (2005). Weka - a machine learning workbench for data mining. Data Mining and Knowledge Discovery Handbook O. Maimon and L. Rokach, Springer-Verlag.

Frankel, D. S., P. Harmon, et al. (2003). The Zachman Framework and the OMG's Model Driven Architecture, Business Process Trends Gartner Group.

Frawley, W., G. Piatetsky-Shapiro, et al. (1992). "Knowledge Discovery in Databases: An Overview." AI Magazine 13(3): 57-70.

Freshmeat.net. (2002). "Freshmeat Software component search engine homepage." Retrieved 26 Jan. 2007, from http://freshmeat.net/.

Gao, J. Z. and Y. Wu (2004). Testing Component-Based Software – Issues, Challenges, and Solutions. 2004 International Conference on COTS-Based Software Systems, Redondo Beach, CA, USA.

Garside, R. and N. Smith (1997). A hybrid grammatical tagger: CLAWS4. Corpus Annotation: Linguistic Information from Computer Text Corpora. R. Garside, G. Leech and A. McEnery. London, Longman: 102-121.

GATE. (2005). "GATE - General Architecture for Text Engineering." Retrieved 16 Sep 2007, from http://gate.ac.uk/.

GFO. (2007). "The General Formal Ontology." Retrieved 9 Feb 2007, from http://www.onto-med.de/en/theories/gfo/index.html.

Gligorov, R., Z. Aleksovski, et al. (2007). Using Google Distance to weight approximate ontology matches. 16th international conference on World Wide Web, Banff Alberta Canada, ACM Press.

Gomez-Perez, A. and O. Corcho (2002). "Ontology languages for the semantic web." IEEE Intelligent Systems 17(1): 54-60.

Google. (2006). "Google search engine homepage." Retrieved 6 Nov 2006, from http://www.google.com.

Google. (2008). "The Official Google Blog." Retrieved 20 Jul 2009, from http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

Gottlob, F. (1980). The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number, Northwestern University Press U.S.

Grace, A. and L. L. Wrage (2004). A case study in COTS product integration using XML. 3rd International Conference on COTS-Based Software Systems, Redondo Beach CA USA, Springer-Verlag

Gracia, J., R. Trillo, et al. (2006). Querying the Web: A Multiontology Disambiguation Method. 6th International Conference on Web Engineering, Palo Alto California USA, ACM Press.

Grau, G., J. P. Carvallo, et al. (2004). DesCOTS: A Software System for Selecting COTS Components. 30th EUROMICRO Conference, Rennes France.

Graubmann, P. and M. Roshchin (2006). Semantic Annotation of Software Components. 32nd EUROMICRO Conference on Software Engineering and Advanced Applications, Cavtat Dubrovnik, IEEE Computer Society Press.

Gregor, S., J. Hutson, et al. (2002). Storyboard Process to Assist in Requirements Verification and Adaptation to Capabilities Inherent in COTS. 1st International Conference on COTS-Based Software Systems, Orlando FL USA, Springer-Verlag.

Grobelnik, M. and D. Mladenic (1998). Learning Machine: design and implementation, Department for Intelligent Systems, J.Stefan Institute.

Gruber, T. R. (1993). "A translation approach to portable ontology specifications." Knowledge Acquisition 5(2): 199 - 220.

Hand, D., H. Mannila, et al. (2001). Principles of Data Mining, MIT Press.

Haykin, S. S. (1994). Neural networks : a comprehensive foundation, Maxwell Macmillan Int.

Hearst, M. A. (2006). "Clustering versus faceted categories for information exploration " Communications of the ACM 49(4): 59-61.

Hepple, M. (2000). Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, Association for Computational Linguistics.

Hiemstra, D. (2001). Using Language Models for Information Retrieval. Enschede, The Netherlands, University of Twente. Ph.D.

Hope, P., G. McGraw, et al. (2004). "Misuse and Abuse Cases: Getting Past the Positive." IEEE Security and Privacy 2(3): 90-92.

Howison, J., M. Conklin, et al. (2004). "FLOSSmole: A Collaborative Repository for FLOSS Research Data and Analyses." International Journal of Information Technology and Web Engineering 1(3): 17-26.

IBM. (2008). "OmniFind." Retrieved 11 Dec. 2008, from http://www-01.ibm.com/.

Ingwersen, P. (1998). "The calculation of web impact factors." Journal of Documentation 54(2): 236-243.

ISO/IEC (2001). "Standard 9126-1 Software Engineering - Product Quality - Part1: Quality Model."

Jain, A. K., M. N. Murty, et al. (1999). "Data clustering: a review." ACM computing surveys 31(3): 264-323.

Jin, Y. and J. Han (2005). Spesifying Interaction Constraints of Software Component for Better Understandability and Interoperability. 4th International Conference on COTS-Based Software Systems, Bilbao Spain, Springer-Verlag.

Johnson, S. C. (1967). "Hierarchical clustering schemes." Psychometrika 32(2): 241-254.

Kapetanios, E., V. Sugumaran, et al. (2008). "A parametric linguistics based approach for cross-lingual web querying." Data and Knowledge Engineering 66(1): 35-52.

Karypis, G., E.-H. Han, et al. (1999). "Chameleon: hierarchical clustering using dynamic modeling." Computer 32(8): 68-75.

Kawaguchi, S., P. Garg, et al. (2006). "MUDABlue: An automatic categorization system for Open Source repositories." Journal of Systems and Software 79(7): 939-953.

Kim, S., H. Alani, et al. (2002). Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. 15th European Conference on Artificial Intelligence, Lyon France.

Kontino, J., G. Caldiera, et al. (1996). Defining factors, goals and criteria for reusable component evaluation. 1996 conference of the Centre for Advanced Studies on Collaborative research, Toronto Ontario Canada, IBM Press.

Koshman, S., A. Spink, et al. (2006). "Web searching on the Vivisimo search engine." Journal of the American Society for Information Science and Technology 57(14): 1875-1887.

Kunda, D. and L. Brooks (2000). "Identifying and Classifying Processes (traditional and soft factors) that Support COTS Component Selection: A Case Study." European Journal of Information Systems 9(4): 226-234.

Kunder, M. d. (2009). "The Indexed Web size estimate." Retrieved 20 Jul 2009, from http://www.worldwidewebsize.com/.

Kural, Y., S. Robertson, et al. (2001). "Deciphering cluster representations." Information Processing and Management 37(4): 593-601.

KW. (2006). "Knowledge Web." Retrieved 17 Sep 2006, from http://knowledgeweb.semanticweb.org.

Landauer, T. and S. Dumais (1997). "A Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological Review 104(2): 211-240.

Laprun, C., J. G. Fiscus, et al. (2002). A practical introduction to ATLAS. 3rd International Conference on Language Resources and Evaluation, Las Palmas Spain.

Larsson, S., I. Crnkovic, et al. (2004). On the Expected Synergies between Component-Based Software Engineering and Best Practices in Product Integration. 30th Euromicro Conference, Rennes France, IEEE Computer Society Press.

Lenneberg, E. (1967). Biological Foundations of Language. New York, John Wiley & Sons, Inc.

Lewis, G. A. and E. J. Morris (2004). From System Requirements to COTS Integration Criteria. 3rd International Conference on COTS-Based Software Systems, Redondo Beach CA USA, Springer-Verlag.

Li, M. and P. Vitanyi (1997). An Introduction to Kolmogorov Complexity and Its Applications. New York, Springer-Verlag.

Liu, T.-Y., Y. Yang, et al. (2005). An experimental study on large-scale web categorization. 14th International Conference on World Wide Web, Chiba Japan, ACM Press.

LSA. (2003). "LSA - Latent Semantic Analysis." Retrieved 23 Jun. 2007, from http://lsa.colorado.edu.

Maarek, Y. S., R. Faginy, et al. (2000). Ephemeral Document Clustering for Web Applications, IBM Research.

Maiden, N. and C. Ncube (1998). "Acquiring COTS Software Selection Requirements." IEEE Software Volume 15(Issue 2): 45-46.

Mann, W. and S. Thompson (1988). "Rhetorical structure theory: Toward a functional theory of text organization." Text 3(8): 243-281.

Manning, C. D. and H. Schütze (1999). Foundations of Statistical Natural Language Processing, MIT Press.

Manning, C. D., P. Raghavan, et al. (2008). Introduction to Information Retrieval, Cambridge University Press.

Mason, O. J. (2006). The Automatic Extraction of Linguistic Information From Text Corpora, The University of Birmingham. Ph.D.

Masterseek. (2008). "Accoona company search engine homepage." Retrieved 10 Dec. 2008, from http://www.accoona.com/.

Mayfield, J. and T. Finin (2004). Information retrieval on the Semantic Web: Integrating inference and retrieval. ACM SIGIR Workshop on the Semantic Web, Toronto Canada.

McGuinness, D. L. and F. v. Harmelen. (2004). "W3C Recommendation. OWL Web Ontology Language Overview." Retrieved 17 Jan. 2005, from http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Mielnik, J.-C., V. Bouthors, et al. (2004). Using eCots portal for sharing information about software products and the Internet and in corporate intranets. 3rd International Conference on COTS-Based Software Systems Redondo Beach CA USA, Springer-Verlag.

Mielnik, J.-C., B. Lang, et al. (2003). eCots platform: an inter-industrial initiative COTS related information sharing. 2nd International Conference on COTS-Based Software Systems, Ottawa Canada, Springer-Verlag.

Mili, A., R. Mili, et al. (1994). Storing and retrieving software components: a refinement based system. 16th International Conference on Software Engineering, Sorrento Italy, IEEE Computer Society Press.

Miller, E. (2004). "The W3C's Semantic Web Activity: An Update." IEEE Intelligent Systems 19(3): 95-96.

Miller, G. A., R. Beckwith, et al. (1990). "Introduction to wordnet: An on-line lexical database." Journal of Lexicography 3(4): 235-244.

Mobasher, B., R. Cooley, et al. (2000). "Automatic personalization based on Web usage mining." Communications of the ACM 43(8): 142-151.

NBC. (2006). "The British National Corpus (NBC)." Retrieved 4 Feb 2006, from http://www.natcorp.ox.ac.uk/.

Ncube, C. and J. C. Dean (2002). The Limitations of Current Decision-Making Techniques in the Procurement of COTS Software Components. 1st International Conference on COTS-Based Software Systems, Orlando FL USA, Springer-Verlag.

Nelson, T. (1960). "Project Xanadu homepage." Retrieved 13 Jan 2007, from http://www.xanadu.net/.

Northcott, M. and M. Vigder (2005). Managing Dependencies Between Software Products. 4th International Conference on COTS-Based Software Systems, Bilbao Spain, Springer-Verlag.

Noy, N. F. and D. L. McGuinness. (2002). "Ontology Development 101: A Guide to Creating Your First Ontology." Retrieved 26 Oct. 2005, from http://protege.stanford.edu/publications/ontology_development/ontology101.html.

Ochs, M., D. Pfahl, et al. (2001). A method for efficient measurement-based COTS assessment and selection method description and evaluation results. 7th International Software Metrics Symposium, London UK, IEEE Computer Society Press.

OOSPICE. (2006). "Software Process Improvement and Capability dEtermination for Object Oriented/component based software development." Retrieved 2 Feb 2006, from http://www.oospice.com/.

Pahl, C. and M. Casey (2003). Ontology Support for Web Service Processes. 9th European Software Engineering Conference, Helsinki Finland, ACM Press.

Peng, X. and B. Choi (2002). Automatic web page classification in a dynamic and hierarchical way. 5th IEEE International Conference on Data Mining, Washington DC, IEEE Computer Society Press.

Penin, T., H. Wang, et al. (2008). Snippet Generation for Semantic Web Search Engines. 3rd Asian Semantic Web Conference, Bangkok, Thailand, Springer-Verlag.

Podgurski, A. and L. Pierce (1992). Behavior Sampling: A Technique For Automated Retrieval Of Reusable Components. 14th International Conference on Software Engineering, Melbourne Australia, ACM Press.

Ponte, J. M. and W. B. Croft (1998). A language modeling approach to information retrieval. 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne Australia, ACM Press.

Powerset. (2008). "Powerset NLP search engine homepage." Retrieved 21 Jan 2009, from http://www.powerset.com/.

Prieto-Diaz, R. (1990). Implementing faceted classification for software reuse. 12th International Conference on Software Engineering, Nice France, IEEE Computer Society Press.

Prieto-Díaz, R. (1990). "Domain analysis: an introduction." ACM SIGSOFT Software Engineering Notes 15(2): 47-54.

Qi, X. and B. D. Davison (2009). "Web page classification: Features and algorithms." ACM Computing Surveys 41(22): Article No. 12.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. 14th International Joint Conference on Artificial Intelligence, Montreal Canada, Lawrence Erlbaum Associates Ltd.

Resnik, P. (1999). "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language." Journal of Artificial Intelligence Research 11: 95-130.

Riley, J. and K. A. Delic (1998). Augmenting Information Retrieval by Knowledge Infusion, Hewlett-Packard.

Robertson, S. E. and K. S. Jones (1976). "Relevance weighting of search terms." Journal of the American Society for Information Science and Technology 27(3): 129-146.

Robertson, S.E. and M.M. Hancock-Beaulieu (1992). “On the evaluation of IR systems.” Information Processing and Management 28(4): 457-466.

Roman, D., U. Keller, et al. (2005). "Web Service Modeling Ontology." Applied Ontology 1(1): 77-106.

Rubenstein, H. and J. B. Goodenough (1965). "Contextual correlates of synonymy." Communications of the ACM 8(10): 627-633.

Sahoo, N., J. Callan, et al. (2006). Incremental hierarchical clustering of text documents. 15th ACM international conference on Information and knowledge management, Arlington Virginia USA, ACM Press.

Salton, G. (1971). The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice-Hall, Inc.

Salton, G., A. Wong, et al. (1975). "A vector space model for automatic indexing." Communications of the ACM 18(11): 613-620.

Salton, G. and M. J. McGill (1983). Introduction to Modern Information Retrieval, McGraw-Hill.

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc.

SEKT. (2007). "SEKT Project." Retrieved 16 Jul 2007, from http://www.sekt-project.com.

Sheth, A. and C. Ramakrishnan (2003). "Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis." IEEE Data Engineering Bulletin 26(4): 40-48.

Singhal, A. (2001). "Modern Information Retrieval: A Brief Overview." IEEE Data Engineering Bulletin 24(4): 35-43.

Sjachyn, M. and L. Beus-Dukic (2006). Semantic component selection - SemaCS. 5th International Conference on COTS-Based Software Systems, Orlando Florida USA, IEEE Computer Society Press.

Song, D. and P. Bruza (2006). Text Based Knowledge Discovery with Information Flow Analysis. 8th Asia Pacific Web Conference, Harbin China, Springer-Verlag.

SourceForge.net. (1999). "SourceForge Software component search engine homepage." Retrieved 10 Oct.2006, from http://SourceForge.net.

Stanford. (2005). "The Protégé Ontology Editor and Knowledge Acquisition System." Retrieved 1 Jan 2007, from http://protege.stanford.edu/.

StatSoft. (2006). "Data Mining Techniques." Retrieved 14 Sep. 2006, from http://www.statsoft.com/textbook/stdatmin.html.

Steels, L. (2007). Language as a Complex Adaptive System. Parallel Problem Solving from Nature PPSN VI, Springer-Verlag. 1917**:** 17-26.

Stefanowski, J. and D. Weiss (2003). Carrot2 and Language Properties in Web Search Results Clustering. 1st International Atlantic Web Intelligence Conference, Madrid Spain, Springer-Verlag.

Stoica, E. and M. A. Hearst (2004). Nearly-automated Metadata Hierarchy Creation. Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting, Companion Volume, Boston USA.

Stoica, E. and M. A. Heart (2006). Demonstration: Using WordNet to Build Hierarchical Facet Categories. International ACM SIGIR Workshop on Faceted Search, Seattle WA USA, ACM Press.

Strohman, T., D. Metzler, et al. (2004). Indri: A language-model based search engine for complex queries, Department of Computer Science, University of Massachusetts.

Sugumaran, V. and V. C. Storey (2003). "A Semantic-Based Approach to Component Retrieval." ACM SIGMIS Database 34(3): 8-24.

SUMO. (2006). "Suggested Upper Merged Ontology (SUMO)." Retrieved 1 Feb 2006, from http://www.ontologyportal.org/.

SUO, (2003). ."Standard Upper Ontology Working Group (SUO)." Retrieved 1 Feb 2006, from http://suo.ieee.org/.

Swoogle. (2007). "Semantic web search engine homepage." Retrieved 10 Dec. 2008, from http://swoogle.umbc.edu/.

Szyperski, C. (2003). Component Technology - What, Where, and How? 25th International Conference on Software Engineering, Portland Oregon, IEEE Computer Society Press.

Szyperski, C., D. Gruntz, et al. (2002). Component software : beyond object-oriented programming, ACM Press.

Torchiano, M., L. Jaccheri, et al. (2002). COTS Products Characterization. 14th International Conference on Software Engineering and Knowledge Engineering, Ischia Italy, ACM Press.

TREC. "Text REtrieval Conference." Retrieved 16/03/2006, from http://trec.nist.gov/.

Uschold, M. and M. Gruninger (2004). "Ontologies and semantics for seamless connectivity." ACM SIGMOD Record 33(4): 58-64.

Vanderlei, T. A., F. A. Durão, et al. (2007). A cooperative classification mechanism for search and retrieval software components. ACM symposium on Applied computing Seoul Korea, ACM Press.

Voas, J. (1998). "COTS software: the economical choice?" IEEE Software 15(2): 16-19.

Voelter, M. (2003). "A Taxonomy for Components." Journal of Object Technology 2(4): 119-125.
Vogel, D., S. Bickel, et al. (2005). "Classifying search engine queries using the web as background knowledge." ACM SIGKDD Explorations Newsletter 7(2): 117-122.
Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh Pennsylvania USA, ACM Press.
Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. 8th Text REtrieval Conference, NIST Special Publication 500-246.
W3C. (2001). "Web Services Description Language (WSDL) 1.1." Retrieved 28 Oct. 2009, from http://www.w3.org/TR/wsdl.
W3C. (2004a). "W3C Recommendation. RDF Primer." Retrieved 22 Dec 2005, from http://www.w3.org/TR/rdf-primer/.
W3C. (2004b). "W3C Recommendation. XML Schema Part 0: Primer Second Edition." Retrieved 23 Dec 2005, from http://www.w3.org/TR/xmlschema-0/.
W3C. (2008). "W3C Recommendation. Extensible Markup Language (XML) 1.0 (Fifth Edition)." Retrieved 23 Dec. 2008, from http://www.w3.org/TR/2008/REC-xml-20081126/.
Wallach, H. M. (2004). Evaluation metrics for hard classifiers, Cavendish Laboratory, University of Cambridge.
Wallnau, K. C., D. Carney, et al. (1998). COTS Software Evaluation, Software Engineering Institute, Carnegie Mellon University.
Wang, P., B.-W. Xu, et al. (2004). A novel Approach to Semantic Annotation Based on Multi-Ontologies. 3rd International Conference on Machine Learning and Cybernetics, Shanghai China, IEEE Computer Society Press.
Welty, C. A. and D. A. Ferrucci (1999). A Formal Ontology for Re-Use of Software Architecture Documents. 14th IEEE International Conference on Automated Software Engineering, Cocoa Beach FL USA, IEEE Computer Society Press.
White, R. W., S. T. Dumais, et al. (2009). Characterizing the Influence of Domain Expertise on Web Search Behavior. 2nd International Conference on Web Search and Data Mining, Barcelona Spain, ACM Press.
White, R. W., G. Muresan, et al. (2006). "Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems." ACM SIGIR Forum 40(2): 52-60.
Whitehead, K. (2002). Component-based development : principles and planning for business systems. London, Addison-Wesley.
Wibowo, W. and H. E. Williams (2002). Strategies for minimising errors in hierarchical web categorisation. 11th International Conference on Information and Knowledge Management, McLean Virginia USA, ACM Press.
Witten, I. H. and E. Frank (2002). "Data mining: practical machine learning tools and techniques with Java implementations." ACM SIGMOD Record 31(1): 76-77.
Witten, I. H. and E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann.
Wolfram, S. (2009). "Wolfram|Alpha answer engine homepage." Retrieved 14 Mar 2009, from http://www.wolframalpha.com/.
Wu, X., V. Kumar, et al. (2007). "Top 10 algorithms in data mining " Knowledge and Information Systems 14(1): 1-37.
Xu, J. and W. B. Croft (1996). Query expansion using local and global document analysis. 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich Switzerland, ACM Press.

Y. Li, K. B. and H. Cunningham (2005). SVM Based Learning System For Information Extraction. Deterministic and Statistical Methods in Machine Learning. J. Winkler, M. Niranjan and N. Lawerence, Springer Verlag. 3635**:** 319-339.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. 33rd annual meeting on Association for Computational Linguistics, Cambridge Massachusetts, Association for Computational Linguistics.

Ye, F. and T. Kelly (2004). COTS - product selection for safety critical systems. 3rd International Conference on COTS-Based Software Systems, Redondo Beach CA USA, Springer-Verlag.

Yeung, C. M. A., N. Gibbins, et al. (2008). Web Search Disambiguation by Collaborative Tagging. The Annual European Conference on Information Retrieval, Glasgow Scotland.

Yigang Xu and P. Degoulet (2001). Using XML in a Component Based Mediation Architecture for the Integration of Applications. XML Europe 2001 Conference, Berlin Germany.

Zaremski, A. M. and J. M. Wing (1995). "Signature Matching: A Tool for Using Software Libraries." Communications of the ACM 4(2): 146-170.

Appendix A: SourceForge.net case study

A1: SourceForge.net case study data

id	name	description
1	Audacity	A fast multi-track audio editor and recorder for Linux, BSD, Mac OS, and Windows. Supports WAV, AIFF, Ogg, and MP3 formats. Features include envelope editing, mixing, built-in effects and plug-ins, all with unlimited undo.
2	VirtualDub	Desktop video processing and capture application (Win32).
3	CDex	CDex is a CD-Ripper, extracting digital audio data from an Audio CD. The application supports many Audio encoders, like MPEG (MP2,MP3), VQF, AAC encoders.
4	AC3Filter	It is DirectShow AC3 Decoder filter used to palyback AVI files with AC3 sound tracks and DVDs. Multichannel and S/PDIF support. Focused at flexible controls during playback: gains, mixer, stream information, levels and other.
5	NASA WorldWind	NASA World Wind is a graphically rich 3D virtual globe for use on desktop computers running Windows. It combines NASA imagery generated from satellites that have produced Blue Marble, Landsat 7, SRTM, MODIS and more.
6	FileZilla	FileZilla is a fast FTP and SFTP client for Windows with a lot of features. FileZilla Server is a reliable FTP server.
7	aMSN	A very nice MSN compatible messenger application, aMSN Messenger is a multiplatform MSN messenger clone. Works pretty much like its Windows based counterpart. Perfect for keeping in touch with those friends who have not yet seen the light. Works on linux
8	UltraVNC	UltraVNC: Remote desktop support software - Remote PC access - remote desktop connection software - VNC Compatibility - FileTransfer - Encryption plugins - Text chat - MS authentication
9	PeerGuardian	PeerGuardian helps protect your privacy by blocking many ranges of aggressive IPs while you use P2P.
10	Wireshark	Wireshark is a powerful network protocol analyzer developed by an international team of networking experts. It runs on UNIX, OS X and Windows. (If you're looking for Ethereal, we switched names in May 2006 due to trademark issues.)
11	aamirplayer	Aamir Media player plays 30 formats of audio/video files with very small size but with great functionality.
12	abcrypt	A C++/Qt program for use in encrypting and decrypting simple substitution cyphers. These cyphers are often found in newspapers and various puzzle books.
13	absinth	Absinth is a C++ Object-Oriented Multiprocess Multithreaded Proxy Server. Able to serve a great number of clients as a little LAN

14	accada-epcis	The objective of the project is to create an implementation of the EPCIS Query and Capture interfaces which allows users to turn their MySQL database into an EPCIS Repository.
15	acdev	Third-Party tools, utilities and resources for use with Turbine's "Asheron's Call" MMORPG
16	aclibico	AC.lib-ICO is a Java 1.4+ library/stand alone program to read image files in MS ICO format (i.e., favicon.ico) from URLs, files, buffers, or streams, and convert them to Image objects (library) or display them (stand alone program).
17	act	ACT (Another Chatting Tool) is a plugin-based chat messanger
18	acudos	AcuDOS is an attempt to revive the PDE MS-DOS Emulator. It aims at emulating the PC hardware and simulating MS-DOS with maximum of accurateness including cycle-exact CPU emulation and support for CGA tweaked modes and true CGA colours.
19	adcviewer	ADCViewer visualizes analog signals coming from a microcontroller board through the serial port.
20	adkp	Aurum WoW Guild DKP system
21	adr	A virtual reality (VR) system for the Internet based on a secure distributed object system. ADR has been renamed and moved: please see new site interreality.org.
22	advisor	Advisor aims to be a comprehensive and widely available, free vulnerability database and management system.
23	aetherion	The aim of the Aetherion project is to develop a MMORPG game. Please see <A HREF="http://aetherion.sf.net/forum/index.php">the forum</A> for more information
24	aftp	aftp is an FTP library and an FTP tool. The aftp tool is a good example on how to use the library. You need FTP functionality in your software? Use the aftp library! Now also atelnet(d), a client and server telnet between MS-Windows (server) and Linux.
25	aggregator	Content Management System
26	aguita	AGUITA -- Apache Graphical User Interface for Total Administration
27	aigo-online	That is an Online Multiplayer Games based on aigochess which is a special chess game in China. It offers Lobby, Chat Rooms, Master Server, Audio/Video sharing play. It's the official traning & game software for China aigo Chess Club,Beijing China.
28	aircraftsched	The AircraftScheduling program is designed to allow the scheduling of aircraft and the operation of an FBO or flying club. The program is used to schedule aircraft and if the administrator chooses to allow the checkin and checkout of aircraft.
29	aixtoolbox	The AIX Toolbox for Linux Applications contains free open source software built for AIX 5L, packaged with RPM. It includes a variety of utilities & libraries often found on Linux distros: http://www-1.ibm.com/servers/aix/products/aixos/linux/index.html

30	ajaxos	AjaxOS is simple ajax framework. Makes possible to easy create insulated application. include modules: xslt, io, user interface, timer(thread), core messages.
31	backslash	backSlash is an open-source text editor for people using several operating systems regularly. Instead of using notepad on Windows, gedit or kedit on linux, they will have only one.
32	elmtprops	ElementProps is an Internet Explorer context menu extension that displays the properties of any DOM node in the current web page, including dynamically created elements and attributes.
33	gflow	Gnome Flow is program written to calculate and visualize simple, steady-state fluid flows around objects. The program calculates the stream, vorticity and pressure using the relaxation method.
34	javapokerserver	Network Card-Poker game written in Java. Server interacts via console with administrator and stores login data in a MySQL database (no GUI planned for server). Client uses an open interface. Default client written in Java as an applet.
35	meta-framework	PHP Meta Framework is a abstract wrapper on different tools to make them work together. Future features: CMS, CMF, MVC, Inversion of control, Design Patterns, Database layer, Forms generation. Main qualities: flexibility, integration, mobility.
36	nxbre	NxBRE is a lightweight Business Rule Engine (aka Rule Based Engine) for the .NET platform, composed of a forward-chaining inference engine and an XML-driven flow control engine. It supports RuleML 0.9 Naf Datalog and Visio 2003 modeling.
37	phpmanta	phpManta is a suite of PHP classes, scripts and examples intended to help PHP programmers writing stable PHP websites and applications. Coding is faster using auto-documentation, templates and web widgets from the phpManta suite.
38	qpoprsp	Project to convert completely original and unique Board-Game design into on-line multiplayer game. Fast paced, high energy. Allows Pure Strategy or Dumb Luck. Many "Joys-of-Winning" and "Agonies-of-Defeat" See more http://www.qpop.com/rsppproject.htm
39	savecracker	SaveCracker is a binary file compare utility that allows users to modify their existing video game save files. SaveCracker users have the ability to replenish their in-game health, items, stats, or even unlock hidden levels and characters.
40	snowmonkey	SnowMonkey is a binding library between C++ and SpiderMonkey (Mozilla's implementation of JavaScript engine). The goal of this project is to create library which will take advantage of boost libraries in order to hide as many binding details as possible
41	TransJVM	TransJVM is a Java package to assist compiler writers targeting the Java Virtual Machine. It provides a simple logical interface for creating JVM classes, expressed in terms of Methods, Fields, Statements and Expressions. JVM details are hidden.
42	Verbum	Verbum is a multilingual dictionary application with extra features to help the user learn the words. It enables you to create your original wordbook, practice & test. There are about 850 words for the Japanese dictionary and 450 for the Italian.

43	wwasher	Using the Linux partition of a dual boot system, this software will provide the capability to eliminate spyware, adware, and other malware from the windows partition. Additional capabilities include the ability to tweak and improve windows from linux.
44	ulogger	Universal Logger is a java server(log4j) + console(lumbermill++) for storing events from log4j, log4cpp, ... in a DB and then use an action appender to forward events to specifics appenders regarding (regexp) the level, category, or message of the event.
45	transitmodel	Transit is an academic project which involves the creation of an XML based scripting language, which allows developers to easily create custom models for long running transactions. An API is also avaliable, allowing easy integration into other projects.
46	threedeemuck	Multi-User Character Kingdom (MUCK) engine and client providing a first person view point 3D-world which can be created by users, and which uses a system to provide quick gameplay by slowly downloading Objects/Textures over time once in the world.
47	tariff-eye	TariffEye is a business intelligence software for analysis of banking tariff. It has a builder to construct and digitalize tariff and a reader to simulate portofolios for further analysis and forecasting of banking costs. May be adapted for other sectors
48	suggest	An easy-to-use JavaScript "class" that adds autocomplete dropdown functionality to HTML text input fields through RPC/Ajax. Similar to Google Suggest, it allows multiple instances per page and is compatible with every major browser except Safari.
49	sqlmap	sqlmap is an automatic SQL injection tool. It is capable to perform an extensive DBMS back-end fingerprint, retrieve remote usernames, tables, columns, enumerate entire DBMS and much more taking advantage of web application SQL injection vulnerabilities.
50	slimey	Slimey is an open source web-based slideshow editor, born as part of the OpenGoo project. Slimey aims to be simple yet provide the most common features you would expect from a slideshow editor. Slimey aims to be compatible with all popular web browsers.
51	scts	The SCTS (Simulated Cable Training System) is a troubleshooting & training system for students that are learning about termination of network cables, using the 568A, 568B, & USOC wiring standards. Multiplayer capabilities are currently being added.

A2: SourceForge.net case study log

SemaCS rules: remove results with score less ||= 3, unless results would be == 1, then remove all results with score less then ||= 2
Experiment run 10 Jul 2008 Data:
SemaSearch:: Scenario NO: 02; Request string: network traffic monitoring tools; Time search started: Thu Jul 10 12:16:18 BST 2008; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Search term monitoring matches T1: servers; Score: 0.9835086749007805; Search term tools matches T3: tools; Score: 0.0; Matched components: ID: 10 Score: 5| ID: 51 Score: 5| ID: 24 Score: 5| ID: 13 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 34 Score: 3| ID: 35 Score: 3| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 36 Score: 1| ID: 15 Score: 1| ; Time results returned: Thu Jul 10 12:18:18 BST 2008; User returned results at: Thu Jul 10 12:19:28 BST 2008; User section: 10,
SemaSearch:: Scenario NO: 04; Request string: video audio player; Time search started: Thu Jul 10 12:22:43 BST 2008; Search term video matches T2: video; Score: 0.0; Search term audio matches T1: audio; Score: 0.0; Search term player matches T3: player; Score: 0.0; Matched components: ID: 3 Score: 14| ID: 1 Score: 10| ID: 11 Score: 8| ID: 27 Score: 4| ID: 4 Score: 3| ID: 39 Score: 2| ID: 2 Score: 2| ID: 21 Score: 2| ID: 33 Score: 1| ID: 48 Score: 1| ID: 16 Score: 1| ID: 8 Score: 1| ID: 12 Score: 1| ID: 18 Score: 1| ID: 47 Score: 1| ID: 32 Score: 1| ; Time results returned: Thu Jul 10 12:23:18 BST 2008; User returned results at: Thu Jul 10 12:25:03 BST 2008; User section: 3, 11, 27, 4,
SemaSearch:: Scenario NO: 01; Request string: ftp upload software; Time search started: Thu Jul 10 12:25:55 BST 2008; Search term ftp matches T2: ftp; Score: 0.0; Search term upload matches T1: image; Score: 0.9439446237013213; Search term software matches T3: software; Score: 0.0; Matched components: ID: 24 Score: 12| ID: 8 Score: 7| ID: 29 Score: 7| ID: 6 Score: 6| ID: 27 Score: 6| ID: 43 Score: 6| ID: 47 Score: 4| ID: 16 Score: 3| ID: 49 Score: 3| ID: 42 Score: 3| ID: 48 Score: 2| ID: 31 Score: 2| ID: 35 Score: 2| ID: 18 Score: 2| ID: 46 Score: 1| ID: 2 Score: 1| ID: 38 Score: 1| ID: 11 Score: 1| ID: 4 Score: 1| ID: 36 Score: 1| ID: 1 Score: 1| ID: 45 Score: 1| ID: 13 Score: 1| ID: 41 Score: 1| ; Time results returned: Thu Jul 10 12:27:14 BST 2008; User returned results at: Thu Jul 10 12:29:17 BST 2008; User section: 24, 6,
SemaSearch:: Scenario NO: 05; Request string: sound conversion; Time search started: Thu Jul 10 12:33:58 BST 2008; Search term sound matches T2: sound; Score: 0.0; Search term conversion matches T2: formats; Score: 1.1659090320390078; Matched components: ID: 1 Score: 5| ID: 4 Score: 3| ID: 11 Score: 3| ; Time results returned: Thu Jul 10 12:34:34 BST 2008; User returned results at: Thu Jul 10 12:35:24 BST 2008; User section: 1,
ERR: Text Search:: Scenario NO: 05; Request string: audio editting; Time search started: Thu Jul 10 12:37:28 BST 2008; ID: 1 Score: 1| ID: 3 Score: 1| ID: 11 Score: 1| ; Time results returned: Thu Jul 10 12:37:29 BST 2008; User returned results at: Thu Jul 10 12:38:18 BST 2008; User section: 1, 3,
ERR: Text Search:: Scenario NO: 01; Request string: file transfer ; Time search started: Thu Jul 10 12:39:18 BST 2008; ID: 4 Score: 1| ID: 11 Score: 1| ID: 16 Score: 1| ID: 39 Score: 1| ; Time results returned: Thu Jul 10 12:39:18 BST 2008; User returned results at: Thu Jul 10 12:40:33 BST 2008; User section:
SemaSearch:: Scenario NO: 03; Request string: software download free remote PC control; Time search started: Fri Jul 11 16:58:43 BST 2008; Search term software matches T3: software; Score: 0.0; Search term download matches T1: http; Score: 0.678429415418877; Search term free matches T3: free; Score: 0.0; Search term remote matches T3: remote; Score: 0.0; Search term PC matches T3: pc; Score: 0.0; Search term control matches T3: control; Score: 0.0; Matched components: ID: 29 Score: 18| ID: 8 Score: 16| ID: 38 Score: 12| ID: 23 Score: 11| ID: 24 Score: 7| ID: 21 Score: 7| ID: 45 Score: 6| ID: 43 Score: 5| ID: 36 Score: 5| ID: 41 Score: 4| ID: 18 Score: 4| ID: 35 Score: 4| ID: 47 Score: 3| ID: 5 Score: 3| ID: 49 Score: 3| ID: 40 Score: 3| ID: 4 Score: 3| ID: 48 Score: 3| ID: 34 Score: 3| ID: 28 Score: 3| ID: 50 Score: 2| ID: 14 Score: 2| ID: 15 Score: 2| ID: 44 Score: 2| ID: 22 Score: 2| ID: 51 Score: 2| ID: 27 Score: 1| ID: 16 Score: 1| ID: 42

Score: 1| ID: 26 Score: 1| ID: 9 Score: 1| ID: 1 Score: 1| ID: 10 Score: 1| ID: 12 Score: 1| ID: 31 Score: 1| ID: 3 Score: 1| ID: 30 Score: 1| ; Time results returned: Fri Jul 11 17:02:08 BST 2008; User returned results at: Fri Jul 11 17:04:49 BST 2008; User section: 8,
Text Search:: Scenario NO: 05; Request string: sound file converter ; Time search started: Fri Jul 11 17:06:18 BST 2008; ID: 4 Score: 2| ID: 6 Score: 1| ID: 8 Score: 1| ID: 11 Score: 1| ID: 16 Score: 1| ID: 39 Score: 1| ; Time results returned: Fri Jul 11 17:06:18 BST 2008; User returned results at: Fri Jul 11 17:08:37 BST 2008; User section: 4, 11,
Text Search:: Scenario NO: 01; Request string: download file server; Time search started: Fri Jul 11 17:09:14 BST 2008; ID: 6 Score: 2| ID: 46 Score: 1| ID: 4 Score: 1| ID: 8 Score: 1| ID: 11 Score: 1| ID: 16 Score: 1| ID: 39 Score: 1| ID: 13 Score: 1| ID: 24 Score: 1| ID: 27 Score: 1| ID: 29 Score: 1| ID: 34 Score: 1| ID: 44 Score: 1| ; Time results returned: Fri Jul 11 17:09:14 BST 2008; User returned results at: Fri Jul 11 17:11:41 BST 2008; User section: 6, 8,
Text Search:: Scenario NO: 02; Request string: network traffic monitor; Time search started: Mon Jul 14 12:26:18 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ; Time results returned: Mon Jul 14 12:26:18 BST 2008; User returned results at: Mon Jul 14 12:27:30 BST 2008; User section: 10,
Text Search:: Scenario NO: 04; Request string: mpeg4; Time search started: Mon Jul 14 12:31:48 BST 2008; Time results returned: Mon Jul 14 12:31:48 BST 2008; User returned results at: Mon Jul 14 12:31:54 BST 2008; User section:
Text Search:: Scenario NO: 03; Request string: remote application to control computer; Time search started: Mon Jul 14 12:55:08 BST 2008; ID: 49 Score: 3| ID: 8 Score: 2| ID: 2 Score: 2| ID: 7 Score: 2| ID: 29 Score: 2| ID: 30 Score: 2| ID: 37 Score: 2| ID: 42 Score: 2| ID: 4 Score: 2| ID: 5 Score: 2| ID: 35 Score: 2| ID: 3 Score: 1| ID: 1 Score: 1| ID: 10 Score: 1| ID: 13 Score: 1| ID: 14 Score: 1| ID: 15 Score: 1| ID: 16 Score: 1| ID: 17 Score: 1| ID: 18 Score: 1| ID: 22 Score: 1| ID: 23 Score: 1| ID: 24 Score: 1| ID: 26 Score: 1| ID: 28 Score: 1| ID: 31 Score: 1| ID: 33 Score: 1| ID: 34 Score: 1| ID: 38 Score: 1| ID: 39 Score: 1| ID: 40 Score: 1| ID: 41 Score: 1| ID: 43 Score: 1| ID: 44 Score: 1| ID: 45 Score: 1| ID: 46 Score: 1| ID: 47 Score: 1| ID: 48 Score: 1| ID: 50 Score: 1| ID: 19 Score: 1| ID: 36 Score: 1| ; Time results returned: Mon Jul 14 12:55:08 BST 2008; User returned results at: Mon Jul 14 13:00:48 BST 2008; User section: 14,
Experiment run 10 Jul 2008 Alternative:
SemaSearch:: Scenario NO: 01; Request string: download file server; Time search started: Tue Jul 22 16:20:50 BST 2008; Search term download matches T1: http; Score: 0.678429415418877; Search term file matches T3: file; Score: 0.0; Search term server matches T2: server; Score: 0.0; Matched components: ID: 38 Score: 12| ID: 23 Score: 11| ID: 21 Score: 7| ID: 45 Score: 6| ID: 29 Score: 5| ID: 34 Score: 5| ID: 24 Score: 5| ID: 48 Score: 4| ID: 41 Score: 4| ID: 39 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 36 Score: 3| ID: 5 Score: 3| ID: 49 Score: 3| ID: 40 Score: 3| ID: 4 Score: 3| ID: 44 Score: 3| ID: 28 Score: 3| ID: 13 Score: 3| ID: 50 Score: 2| ID: 14 Score: 2| ID: 15 Score: 2| ID: 18 Score: 2| ID: 51 Score: 2| ID: 16 Score: 1| ID: 42 Score: 1| ID: 22 Score: 1| ID: 26 Score: 1| ID: 9 Score: 1| ID: 1 Score: 1| ID: 10 Score: 1| ID: 12 Score: 1| ID: 35 Score: 1| ID: 31 Score: 1| ID: 3 Score: 1| ID: 47 Score: 1| ID: 30 Score: 1| ID: 27 Score: 1| ; Time results returned: Tue Jul 22 16:22:06 BST 2008; User returned results at: Tue Jul 22 16:33:02 BST 2008; User section: 24, 6, 8,
Text Search:: Scenario NO: 01; Request string: ftp upload software; Time search started: Tue Jul 22 16:46:24 BST 2008; ID: 24 Score: 2| ID: 6 Score: 1| ID: 8 Score: 1| ID: 27 Score: 1| ID: 29 Score: 1| ID: 43 Score: 1| ID: 47 Score: 1| ; Time results returned: Tue Jul 22 16:46:24 BST 2008; User returned results at: Tue Jul 22 16:46:35 BST 2008; User section: 24, 6, 8,
Text Search:: Scenario NO: 02; Request string: network traffic monitoring tools; Time search started: Tue Jul 22 16:49:25 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ID: 15 Score: 1| ID: 35 Score: 1| ; Time results returned: Tue Jul 22 16:49:25 BST 2008; User returned results at: Tue Jul 22 16:49:36 BST 2008; User section: 10,
SemaSearch:: Scenario NO: 02; Request string: network traffic monitor; Time search started: Tue Jul 22 16:53:33 BST 2008; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Search term monitor matches T1: servers; Score:

0.6294424386323639; Matched components: ID: 10 Score: 5| ID: 51 Score: 5| ID: 24 Score: 5| ID: 13 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 34 Score: 3| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 36 Score: 1| ; Time results returned: Tue Jul 22 16:54:45 BST 2008; User returned results at: Tue Jul 22 16:55:09 BST 2008; User section: 10,
Text Search:: Scenario NO: 03; Request string: software download free remote PC control; Time search started: Tue Jul 22 17:00:22 BST 2008; ID: 8 Score: 3| ID: 29 Score: 2| ID: 24 Score: 1| ID: 27 Score: 1| ID: 43 Score: 1| ID: 47 Score: 1| ID: 46 Score: 1| ID: 22 Score: 1| ID: 49 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 48 Score: 1| ID: 4 Score: 1| ID: 19 Score: 1| ID: 35 Score: 1| ID: 36 Score: 1| ; Time results returned: Tue Jul 22 17:00:22 BST 2008; User returned results at: Tue Jul 22 17:00:37 BST 2008; User section: 8,
SemaSearch:: Scenario NO: 03; Request string: remote application to control computer; Time search started: Tue Jul 22 17:04:30 BST 2008; Search term remote matches T3: remote; Score: 0.0; Search term application matches T3: application; Score: 0.0; Search term to matches T3: io; Score: 0.05525913810672244; Search term control matches T3: control; Score: 0.0; Search term computer matches T3: mp; Score: 0.5343298238530242; Matched components: ID: 30 Score: 8| ID: 3 Score: 5| ID: 7 Score: 5| ID: 8 Score: 4| ID: 49 Score: 4| ID: 35 Score: 3| ID: 2 Score: 2| ID: 36 Score: 2| ID: 42 Score: 1| ID: 1 Score: 1| ; Time results returned: Tue Jul 22 17:08:00 BST 2008; User returned results at: Tue Jul 22 17:09:26 BST 2008; User section: 8,
Text Search:: Scenario NO: 04; Request string: video audio player; Time search started: Tue Jul 22 17:18:15 BST 2008; ID: 11 Score: 3| ID: 27 Score: 3| ID: 2 Score: 1| ID: 39 Score: 1| ID: 1 Score: 1| ID: 3 Score: 1| ID: 38 Score: 1| ID: 51 Score: 1| ; Time results returned: Tue Jul 22 17:18:15 BST 2008; User returned results at: Tue Jul 22 17:18:25 BST 2008; User section: 11, 1, 3,
SemaSearch:: Scenario NO: 04; Request string: mpeg4; Time search started: Tue Jul 22 17:22:40 BST 2008; Search term mpeg4 matches T3: application; Score: 1.4843261672455363; Matched components: ID: 7 Score: 5| ID: 3 Score: 3| ID: 49 Score: 3| ID: 2 Score: 2| ID: 30 Score: 2| ID: 42 Score: 1| ; Time results returned: Tue Jul 22 17:23:31 BST 2008; User returned results at: Tue Jul 22 17:24:07 BST 2008; User section: 3,
Text Search:: Scenario NO: 05; Request string: sound conversion; Time search started: Tue Jul 22 17:30:54 BST 2008; ID: 4 Score: 1| ; Time results returned: Tue Jul 22 17:30:54 BST 2008; User returned results at: Tue Jul 22 17:31:04 BST 2008; User section:
SemaSearch:: Scenario NO: 05; Request string: sound file converter ; Time search started: Tue Jul 22 17:33:31 BST 2008; Search term sound matches T2: sound; Score: 0.0; Search term file matches T3: file; Score: 0.0; Search term converter matches T2: formats; Score: 1.5196788319220156; Matched components: ID: 1 Score: 5| ID: 39 Score: 4| ID: 4 Score: 3| ID: 11 Score: 3| ; Time results returned: Tue Jul 22 17:34:41 BST 2008; User returned results at: Tue Jul 22 17:34:56 BST 2008; User section: 1,
SemaSearch:: Scenario NO: 03; Request string: remote login ; Time search started: Fri Aug 08 12:20:31 BST 2008; Search term remote matches T3: remote; Score: 0.0; Search term login matches T2: login; Score: 0.0; Matched components: ID: 8 Score: 4| ID: 34 Score: 4| ID: 49 Score: 1| ; Time results returned: Fri Aug 08 12:21:15 BST 2008; User returned results at: Fri Aug 08 12:23:37 BST 2008; User section: 8,
Experiment run 14 Aug 2008 Data:
SemaSearch:: Scenario NO: 04; Request string: audio ; Time search started: Thu Aug 14 11:14:56 BST 2008; Search term audio matches T1: audio; Score: 0.0; Matched components: ID: 3 Score: 7| ID: 1 Score: 5| ID: 11 Score: 3| ID: 4 Score: 3| ID: 27 Score: 2| ID: 2 Score: 1| ID: 39 Score: 1| ID: 33 Score: 1| ID: 48 Score: 1| ID: 16 Score: 1| ID: 8 Score: 1| ID: 12 Score: 1| ID: 21 Score: 1| ID: 18 Score: 1| ID: 47 Score: 1| ID: 32 Score: 1| ; Time results returned: Thu Aug 14 11:15:05 BST 2008; User returned results at: Thu Aug 14 11:17:17 BST 2008; User section: 3, 1, 4, 39, 33, 16, 8, 21, 32,
SemaSearch:: Scenario NO: 05; Request string: sound file conversion; Time search started: Thu Aug 14 11:56:03 BST 2008; Search term sound matches T2: sound; Score: 0.0; Search term file matches T3: file; Score: 0.0; Search term conversion matches T2: formats; Score: 1.1659090320390078; Matched

components: ID: 1 Score: 5| ID: 39 Score: 4| ID: 4 Score: 3| ID: 11 Score: 3| ; Time results returned: Thu Aug 14 11:57:14 BST 2008; User returned results at: Thu Aug 14 12:12:51 BST 2008; User section: 1,
SemaSearch:: Scenario NO: 01; Request string: software to upload files; Time search started: Thu Aug 14 12:05:50 BST 2008; Search term software matches T3: software; Score: 0.0; Search term to matches T3: io; Score: 0.05525913810672244; Search term upload matches T1: image; Score: 0.9439446237013213; Search term files matches T3: files; Score: 0.0; Matched components: ID: 16 Score: 12| ID: 8 Score: 7| ID: 24 Score: 7| ID: 29 Score: 7| ID: 4 Score: 7| ID: 27 Score: 6| ID: 43 Score: 6| ID: 30 Score: 6| ID: 47 Score: 4| ID: 39 Score: 4| ID: 49 Score: 3| ID: 42 Score: 3| ID: 48 Score: 2| ID: 31 Score: 2| ID: 11 Score: 2| ID: 6 Score: 2| ID: 35 Score: 2| ID: 18 Score: 2| ID: 46 Score: 1| ID: 2 Score: 1| ID: 38 Score: 1| ID: 36 Score: 1| ID: 1 Score: 1| ID: 45 Score: 1| ID: 13 Score: 1| ID: 41 Score: 1| ; Time results returned: Thu Aug 14 12:08:31 BST 2008; User returned results at: Thu Aug 14 12:30:58 BST 2008; User section: 24, 6,
SemaSearch:: Scenario NO: 02; Request string: network traffic monitor tool; Time search started: Thu Aug 14 12:38:43 BST 2008; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Search term monitor matches T1: servers; Score: 0.6294424386323639; Search term tool matches T3: tool; Score: 0.0; Matched components: ID: 24 Score: 9| ID: 49 Score: 8| ID: 10 Score: 5| ID: 51 Score: 5| ID: 13 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 34 Score: 3| ID: 17 Score: 2| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 36 Score: 1| ; Time results returned: Thu Aug 14 12:40:30 BST 2008; User returned results at: Thu Aug 14 12:43:11 BST 2008; User section: 10,
SemaSearch:: Scenario NO: 03; Request string: remote control; Time search started: Thu Aug 14 13:09:56 BST 2008; Search term remote matches T3: remote; Score: 0.0; Search term control matches T3: control; Score: 0.0; Matched components: ID: 8 Score: 4| ID: 35 Score: 3| ID: 36 Score: 2| ID: 49 Score: 1| ; Time results returned: Thu Aug 14 13:10:52 BST 2008; User returned results at: Thu Aug 14 13:11:59 BST 2008; User section: 8,
Text Search:: Scenario NO: 01; Request string: application dowload upload server; Time search started: Thu Aug 14 16:01:25 BST 2008; ID: 29 Score: 2| ID: 2 Score: 1| ID: 3 Score: 1| ID: 7 Score: 1| ID: 30 Score: 1| ID: 37 Score: 1| ID: 42 Score: 1| ID: 49 Score: 1| ID: 6 Score: 1| ID: 13 Score: 1| ID: 24 Score: 1| ID: 27 Score: 1| ID: 34 Score: 1| ID: 44 Score: 1| ; Time results returned: Thu Aug 14 16:01:25 BST 2008; User returned results at: Thu Aug 14 16:03:23 BST 2008; User section: 29, 6,
Text Search:: Scenario NO: 02; Request string: monitor network traffic; Time search started: Thu Aug 14 16:12:16 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ; Time results returned: Thu Aug 14 16:12:16 BST 2008; User returned results at: Thu Aug 14 16:12:58 BST 2008; User section: 10, 51,
Text Search:: Scenario NO: 03; Request string: application for remotly control computer; Time search started: Thu Aug 14 16:14:50 BST 2008; ID: 7 Score: 2| ID: 29 Score: 2| ID: 42 Score: 2| ID: 49 Score: 2| ID: 4 Score: 2| ID: 5 Score: 2| ID: 35 Score: 2| ID: 36 Score: 2| ID: 2 Score: 1| ID: 3 Score: 1| ID: 30 Score: 1| ID: 37 Score: 1| ID: 1 Score: 1| ID: 6 Score: 1| ID: 10 Score: 1| ID: 11 Score: 1| ID: 12 Score: 1| ID: 15 Score: 1| ID: 16 Score: 1| ID: 18 Score: 1| ID: 21 Score: 1| ID: 23 Score: 1| ID: 26 Score: 1| ID: 27 Score: 1| ID: 31 Score: 1| ID: 34 Score: 1| ID: 41 Score: 1| ID: 44 Score: 1| ID: 45 Score: 1| ID: 47 Score: 1| ID: 51 Score: 1| ID: 19 Score: 1| ; Time results returned: Thu Aug 14 16:14:50 BST 2008; User returned results at: Thu Aug 14 16:18:37 BST 2008; User section: 49,
Text Search:: Scenario NO: 04; Request string: media player; Time search started: Thu Aug 14 16:19:29 BST 2008; ID: 11 Score: 2| ID: 27 Score: 1| ID: 38 Score: 1| ID: 51 Score: 1| ; Time results returned: Thu Aug 14 16:19:29 BST 2008; User returned results at: Thu Aug 14 16:20:12 BST 2008; User section: 11,
Text Search:: Scenario NO: 05; Request string: sound convertor; Time search started: Thu Aug 14 16:24:06 BST 2008; ID: 4 Score: 1| ; Time results returned: Thu Aug 14 16:24:06 BST 2008; User returned results at: Thu Aug 14 16:24:29 BST 2008; User section: 4,
SemaSearch:: Scenario NO: 01; Request string: file upload; Time search started: Thu Aug 14 16:41:20 BST 2008; Search term file matches T3: file; Score: 0.0; Search term upload matches T1: image; Score: 0.9439446237013213; Matched components: ID: 27 Score: 5| ID: 39 Score: 4| ID: 16 Score: 3| ID: 49 Score: 3|

ID: 42 Score: 3| ID: 48 Score: 2| ID: 31 Score: 2| ID: 6 Score: 2| ID: 47 Score: 2| ID: 35 Score: 2| ID: 18 Score: 2| ID: 46 Score: 1| ID: 2 Score: 1| ID: 8 Score: 1| ID: 38 Score: 1| ID: 11 Score: 1| ID: 4 Score: 1| ID: 36 Score: 1| ID: 43 Score: 1| ID: 1 Score: 1| ID: 45 Score: 1| ID: 13 Score: 1| ID: 41 Score: 1| ; Time results returned: Thu Aug 14 16:42:26 BST 2008; User returned results at: Thu Aug 14 16:45:11 BST 2008; User section: 6,
SemaSearch:: Scenario NO: 02; Request string: network traffic; Time search started: Thu Aug 14 16:51:46 BST 2008; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Matched components: ID: 51 Score: 5| ID: 10 Score: 4| ID: 34 Score: 1| ID: 13 Score: 1| ; Time results returned: Thu Aug 14 16:52:35 BST 2008; User returned results at: Thu Aug 14 16:53:22 BST 2008; User section: 10, 34, 13,
SemaSearch:: Scenario NO: 03; Request string: remote desktopping; Time search started: Thu Aug 14 16:56:22 BST 2008; Search term remote matches T3: remote; Score: 0.0; Search term desktopping matches T0: ; Score: 0.0; Matched components: ID: 8 Score: 4| ID: 49 Score: 1| ; Time results returned: Thu Aug 14 16:57:02 BST 2008; User returned results at: Thu Aug 14 16:57:31 BST 2008; User section: 8,
SemaSearch:: Scenario NO: 04; Request string: mpeg player; Time search started: Thu Aug 14 17:03:05 BST 2008; Search term mpeg matches T3: mpeg; Score: 0.0; Search term player matches T3: player; Score: 0.0; Matched components: ID: 3 Score: 7| ID: 11 Score: 2| ; Time results returned: Thu Aug 14 17:04:01 BST 2008; User returned results at: Thu Aug 14 17:06:35 BST 2008; User section: 3, 11,
SemaSearch:: Scenario NO: 05; Request string: mp3 maker; Time search started: Thu Aug 14 17:11:16 BST 2008; Search term mp3 matches T2: ogg; Score: 0.4025640999929069; Search term maker matches T1: servers; Score: 3.951767875827487; Matched components: ID: 3 Score: 7| ID: 1 Score: 5| ID: 24 Score: 5| ID: 6 Score: 4| ID: 8 Score: 4| ID: 13 Score: 3| ID: 34 Score: 2| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 10 Score: 1| ID: 36 Score: 1| ; Time results returned: Thu Aug 14 17:11:55 BST 2008; User returned results at: Thu Aug 14 17:12:49 BST 2008; User section: 3, 1, 24,
Experiment run 15 Aug 2008 Data:
SemaSearch:: Scenario NO: 01; Request string: ftp client; Time search started: Fri Aug 15 12:08:22 BST 2008; Search term ftp matches T2: ftp; Score: 0.0; Search term client matches T3: client; Score: 0.0; Matched components: ID: 24 Score: 12| ID: 6 Score: 6| ID: 34 Score: 5| ID: 46 Score: 1| ; Time results returned: Fri Aug 15 12:09:21 BST 2008; User returned results at: Fri Aug 15 12:09:31 BST 2008; User section: 6,
Text Search:: Scenario NO: 02; Request string: netwrok traffic; Time search started: Fri Aug 15 13:07:00 BST 2008; Time results returned: Fri Aug 15 13:07:00 BST 2008; User returned results at: Fri Aug 15 13:07:14 BST 2008; User section:
Text Search:: Scenario NO: 03; Request string: remote control computer; Time search started: Fri Aug 15 13:22:15 BST 2008; ID: 8 Score: 1| ID: 49 Score: 1| ID: 4 Score: 1| ID: 19 Score: 1| ID: 35 Score: 1| ID: 36 Score: 1| ID: 5 Score: 1| ; Time results returned: Fri Aug 15 13:22:15 BST 2008; User returned results at: Fri Aug 15 13:23:20 BST 2008; User section: 8,
Text Search:: Scenario NO: 04; Request string: media applications; Time search started: Fri Aug 15 13:53:09 BST 2008; ID: 11 Score: 1| ID: 29 Score: 1| ID: 37 Score: 1| ; Time results returned: Fri Aug 15 13:53:09 BST 2008; User returned results at: Fri Aug 15 13:53:56 BST 2008; User section: 11,
Experiment run 16 Aug 2008 Data:
Text Search:: Scenario NO: 05; Request string: modify sound files; Time search started: Sat Aug 16 13:41:34 BST 2008; ID: 39 Score: 2| ID: 4 Score: 2| ID: 11 Score: 1| ID: 16 Score: 1| ; Time results returned: Sat Aug 16 13:41:34 BST 2008; User returned results at: Sat Aug 16 13:42:37 BST 2008; User section:
1)>>>>query below - user has entered an extremely large number of spaces/tabs (this was not noticed during experiment and has caused a buffer overflow - this overflow was not completely handled in the prototype casing false matches)

Text Search:: Scenario NO: 01; Request string: download onto server ; Time search started: Sat Aug 16 13:55:30 BST 2008; ID: 6 Score: 421| ID: 13 Score: 421| ID: 24 Score: 421| ID: 27 Score: 421| ID: 29 Score: 421| ID: 34 Score: 421| ID: 44 Score: 421| ID: 46 Score: 421| ID: 1 Score: 420| ID: 2 Score: 420| ID: 3 Score: 420| ID: 4 Score: 420| ID: 5 Score: 420| ID: 7 Score: 420| ID: 8 Score: 420| ID: 9 Score: 420| ID: 10 Score: 420| ID: 11 Score: 420| ID: 12 Score: 420| ID: 14 Score: 420| ID: 15 Score: 420| ID: 16 Score: 420| ID: 17 Score: 420| ID: 18 Score: 420| ID: 19 Score: 420| ID: 20 Score: 420| ID: 21 Score: 420| ID: 22 Score: 420| ID: 23 Score: 420| ID: 25 Score: 420| ID: 26 Score: 420| ID: 28 Score: 420| ID: 30 Score: 420| ID: 31 Score: 420| ID: 32 Score: 420| ID: 33 Score: 420| ID: 35 Score: 420| ID: 36 Score: 420| ID: 37 Score: 420| ID: 38 Score: 420| ID: 39 Score: 420| ID: 40 Score: 420| ID: 41 Score: 420| ID: 42 Score: 420| ID: 43 Score: 420| ID: 45 Score: 420| ID: 47 Score: 420| ID: 48 Score: 420| ID: 49 Score: 420| ID: 50 Score: 420| ID: 51 Score: 420| ; Time results returned: Sat Aug 16 13:55:33 BST 2008; User returned results at: Sat Aug 16 13:58:45 BST 2008; User section: 6, 24, 1, 3,
SemaSearch:: Scenario NO: 02; Request string: monitor network traffic; Time search started: Sat Aug 16 14:02:18 BST 2008; Search term monitor matches T1: servers; Score: 0.6294424386323639; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Matched components: ID: 24 Score: 5| ID: 10 Score: 5| ID: 51 Score: 5| ID: 13 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 34 Score: 3| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 36 Score: 1| ; Time results returned: Sat Aug 16 14:03:32 BST 2008; User returned results at: Sat Aug 16 14:07:19 BST 2008; User section: 10, 13, 8, 44, 48,
SemaSearch:: Scenario NO: 03; Request string: controling computer; Time search started: Sat Aug 16 14:08:31 BST 2008; Search term controling matches T3: os; Score: 1.4037456354365745; Search term computer matches T3: mp; Score: 0.5343298238530242; Matched components: ID: 1 Score: 6| ID: 10 Score: 3| ID: 3 Score: 2| ; Time results returned: Sat Aug 16 14:09:53 BST 2008; User returned results at: Sat Aug 16 14:10:42 BST 2008; User section:
Experiment run 18 Aug 2008 Data:
SemaSearch:: Scenario NO: 04; Request string: video playback and audio playback; Time search started: Mon Aug 18 12:50:53 BST 2008; Search term video matches T2: video; Score: 0.0; Search term playback matches T2: playback; Score: 0.0; Search term and matches T3: io; Score: 0.05070387634032431; Search term audio matches T1: audio; Score: 2.262178096311623; Search term playback matches T2: playback; Score: 0.0; Matched components: ID: 3 Score: 14| ID: 1 Score: 10| ID: 4 Score: 9| ID: 11 Score: 6| ID: 30 Score: 6| ID: 27 Score: 4| ID: 39 Score: 2| ID: 2 Score: 2| ID: 21 Score: 2| ID: 33 Score: 1| ID: 48 Score: 1| ID: 16 Score: 1| ID: 8 Score: 1| ID: 12 Score: 1| ID: 18 Score: 1| ID: 47 Score: 1| ID: 32 Score: 1| ; Time results returned: Mon Aug 18 12:52:56 BST 2008; User returned results at: Mon Aug 18 12:57:23 BST 2008; User section: 4, 11, 16,
SemaSearch:: Scenario NO: 05; Request string: mp3 conversastion; Time search started: Mon Aug 18 13:35:05 BST 2008; Search term mp3 matches T2: ogg; Score: 0.4025640999929069; Search term conversastion matches T0: ; Score: 0.0; Matched components: ID: 3 Score: 7| ID: 1 Score: 5| ; Time results returned: Mon Aug 18 13:35:48 BST 2008; User returned results at: Mon Aug 18 13:36:40 BST 2008; User section: 1,
Text Search:: Scenario NO: 01; Request string: ssh secure shell; Time search started: Mon Aug 18 13:55:44 BST 2008; ID: 21 Score: 1| ; Time results returned: Mon Aug 18 13:55:44 BST 2008; User returned results at: Mon Aug 18 13:56:16 BST 2008; User section: 21,
Text Search:: Scenario NO: 02; Request string: monitor network traffic; Time search started: Mon Aug 18 14:00:18 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ; Time results returned: Mon Aug 18 14:00:18 BST 2008; User returned results at: Mon Aug 18 14:01:34 BST 2008; User section: 10,
Text Search:: Scenario NO: 03; Request string: remotly control computer; Time search started: Mon Aug 18 14:31:03 BST 2008; ID: 4 Score: 1| ID: 19 Score: 1| ID: 35 Score: 1| ID: 36 Score: 1| ID: 5 Score: 1| ; Time results returned: Mon Aug 18 14:31:03 BST 2008; User returned results at: Mon Aug 18 14:32:00 BST 2008; User section:

Text Search:: Scenario NO: 04; Request string: media player to plsy video/audio; Time search started: Mon Aug 18 14:45:20 BST 2008; ID: 11 Score: 2| ID: 38 Score: 2| ID: 27 Score: 1| ID: 51 Score: 1| ID: 1 Score: 1| ID: 2 Score: 1| ID: 4 Score: 1| ID: 5 Score: 1| ID: 7 Score: 1| ID: 8 Score: 1| ID: 10 Score: 1| ID: 13 Score: 1| ID: 14 Score: 1| ID: 15 Score: 1| ID: 16 Score: 1| ID: 17 Score: 1| ID: 18 Score: 1| ID: 22 Score: 1| ID: 23 Score: 1| ID: 24 Score: 1| ID: 26 Score: 1| ID: 28 Score: 1| ID: 29 Score: 1| ID: 30 Score: 1| ID: 31 Score: 1| ID: 33 Score: 1| ID: 34 Score: 1| ID: 35 Score: 1| ID: 37 Score: 1| ID: 39 Score: 1| ID: 40 Score: 1| ID: 41 Score: 1| ID: 42 Score: 1| ID: 43 Score: 1| ID: 44 Score: 1| ID: 45 Score: 1| ID: 46 Score: 1| ID: 47 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 50 Score: 1| ; Time results returned: Mon Aug 18 14:45:21 BST 2008; User returned results at: Mon Aug 18 14:47:11 BST 2008; User section: 11, 1,
Text Search:: Scenario NO: 05; Request string: sound file modifier and converter; Time search started: Mon Aug 18 15:11:11 BST 2008; ID: 4 Score: 3| ID: 6 Score: 2| ID: 16 Score: 2| ID: 39 Score: 2| ID: 8 Score: 1| ID: 11 Score: 1| ID: 1 Score: 1| ID: 2 Score: 1| ID: 5 Score: 1| ID: 10 Score: 1| ID: 12 Score: 1| ID: 14 Score: 1| ID: 15 Score: 1| ID: 18 Score: 1| ID: 21 Score: 1| ID: 22 Score: 1| ID: 24 Score: 1| ID: 28 Score: 1| ID: 32 Score: 1| ID: 33 Score: 1| ID: 34 Score: 1| ID: 36 Score: 1| ID: 37 Score: 1| ID: 38 Score: 1| ID: 40 Score: 1| ID: 41 Score: 1| ID: 42 Score: 1| ID: 43 Score: 1| ID: 44 Score: 1| ID: 46 Score: 1| ID: 47 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 51 Score: 1| ; Time results returned: Mon Aug 18 15:11:11 BST 2008; User returned results at: Mon Aug 18 15:13:26 BST 2008; User section: 11, 1,
Experiment run 14 Aug 2008 Alternative:
Text Search:: Scenario NO: 01; Request string: software to upload files; Time search started: Wed Aug 27 15:37:20 BST 2008; ID: 8 Score: 2| ID: 24 Score: 2| ID: 29 Score: 2| ID: 43 Score: 2| ID: 47 Score: 2| ID: 4 Score: 2| ID: 16 Score: 2| ID: 39 Score: 2| ID: 27 Score: 1| ID: 1 Score: 1| ID: 2 Score: 1| ID: 5 Score: 1| ID: 7 Score: 1| ID: 10 Score: 1| ID: 13 Score: 1| ID: 14 Score: 1| ID: 15 Score: 1| ID: 17 Score: 1| ID: 18 Score: 1| ID: 22 Score: 1| ID: 23 Score: 1| ID: 26 Score: 1| ID: 28 Score: 1| ID: 30 Score: 1| ID: 31 Score: 1| ID: 33 Score: 1| ID: 34 Score: 1| ID: 35 Score: 1| ID: 37 Score: 1| ID: 38 Score: 1| ID: 40 Score: 1| ID: 41 Score: 1| ID: 42 Score: 1| ID: 44 Score: 1| ID: 45 Score: 1| ID: 46 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 50 Score: 1| ID: 11 Score: 1| ; Time results returned: Wed Aug 27 15:37:20 BST 2008; User returned results at: Wed Aug 27 15:37:31 BST 2008; User section: 24,
Text Search:: Scenario NO: 02; Request string: network traffic monitor tool; Time search started: Wed Aug 27 15:38:05 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ID: 15 Score: 1| ID: 17 Score: 1| ID: 24 Score: 1| ID: 29 Score: 1| ID: 35 Score: 1| ID: 49 Score: 1| ; Time results returned: Wed Aug 27 15:38:05 BST 2008; User returned results at: Wed Aug 27 15:38:08 BST 2008; User section: 10,
Text Search:: Scenario NO: 03; Request string: remote control; Time search started: Wed Aug 27 15:38:30 BST 2008; ID: 8 Score: 1| ID: 49 Score: 1| ID: 4 Score: 1| ID: 19 Score: 1| ID: 35 Score: 1| ID: 36 Score: 1| ; Time results returned: Wed Aug 27 15:38:30 BST 2008; User returned results at: Wed Aug 27 15:38:33 BST 2008; User section: 8,
Text Search:: Scenario NO: 04; Request string: audio; Time search started: Wed Aug 27 15:38:53 BST 2008; ID: 1 Score: 1| ID: 3 Score: 1| ID: 11 Score: 1| ID: 27 Score: 1| ; Time results returned: Wed Aug 27 15:38:53 BST 2008; User returned results at: Wed Aug 27 15:39:29 BST 2008; User section: 1, 3, 11, 27,
Text Search:: Scenario NO: 05; Request string: sound file conversion; Time search started: Wed Aug 27 15:40:27 BST 2008; ID: 4 Score: 2| ID: 6 Score: 1| ID: 8 Score: 1| ID: 11 Score: 1| ID: 16 Score: 1| ID: 39 Score: 1| ; Time results returned: Wed Aug 27 15:40:27 BST 2008; User returned results at: Wed Aug 27 15:40:30 BST 2008; User section: 4,
SemaSearch:: Scenario NO: 01; Request string: application dowload upload server; Time search started: Wed Aug 27 16:17:43 BST 2008; Search term application matches T3: application; Score: 0.0; Search term dowload matches T2: play; Score: 1.1776900992482322; Search term upload matches T1: image; Score: 0.9439446237013213; Search term server matches T2: server; Score: 0.0; Matched components: ID: 27 Score: 12| ID: 49 Score: 7| ID: 46 Score: 7| ID: 6 Score: 6| ID: 3 Score: 5| ID: 7 Score: 5| ID: 42 Score: 5| ID: 35 Score: 5| ID: 51 Score: 5| ID: 48 Score: 5| ID: 8 Score: 5| ID: 24 Score: 5| ID: 23 Score: 4| ID: 4

Score: 4| ID: 38 Score: 4| ID: 10 Score: 4| ID: 13 Score: 4| ID: 2 Score: 3| ID: 11 Score: 3| ID: 21 Score: 3| ID: 43 Score: 3| ID: 45 Score: 3| ID: 47 Score: 3| ID: 34 Score: 3| ID: 36 Score: 3| ID: 16 Score: 3| ID: 30 Score: 2| ID: 1 Score: 2| ID: 29 Score: 2| ID: 31 Score: 2| ID: 18 Score: 2| ID: 5 Score: 1| ID: 41 Score: 1| ID: 44 Score: 1| ; Time results returned: Wed Aug 27 16:19:46 BST 2008; User returned results at: Wed Aug 27 16:42:11 BST 2008; User section: 6, 8, 24,
SemaSearch:: Scenario NO: 02; Request string: monitor network traffic; Time search started: Wed Aug 27 16:42:27 BST 2008; Search term monitor matches T1: servers; Score: 0.6294424386323639; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Matched components: ID: 24 Score: 5| ID: 10 Score: 5| ID: 51 Score: 5| ID: 13 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 34 Score: 3| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 36 Score: 1| ; Time results returned: Wed Aug 27 16:43:39 BST 2008; User returned results at: Wed Aug 27 16:43:54 BST 2008; User section: 10,
SemaSearch:: Scenario NO: 03; Request string: application for remotly control computer; Time search started: Wed Aug 27 16:44:10 BST 2008; Search term application matches T3: application; Score: 0.0; Search term for matches T3: ac; Score: 0.10265455541375827; Search term remotly matches T2: text; Score: 1.3974317037787685; Search term control matches T3: control; Score: 0.0; Search term computer matches T3: mp; Score: 0.5343298238530242; Matched components: ID: 16 Score: 9| ID: 4 Score: 6| ID: 3 Score: 5| ID: 7 Score: 5| ID: 49 Score: 3| ID: 35 Score: 3| ID: 2 Score: 2| ID: 30 Score: 2| ID: 31 Score: 2| ID: 48 Score: 2| ID: 36 Score: 2| ID: 42 Score: 1| ID: 8 Score: 1| ID: 46 Score: 1| ID: 38 Score: 1| ID: 1 Score: 1| ; Time results returned: Wed Aug 27 16:47:15 BST 2008; User returned results at: Wed Aug 27 16:47:33 BST 2008; User section: 8,
SemaSearch:: Scenario NO: 04; Request string: media player; Time search started: Wed Aug 27 16:47:58 BST 2008; Search term media matches T3: media; Score: 0.0; Search term player matches T3: player; Score: 0.0; Matched components: ID: 11 Score: 4| ; Time results returned: Wed Aug 27 16:49:10 BST 2008; User returned results at: Wed Aug 27 16:49:21 BST 2008; User section: 11,
SemaSearch:: Scenario NO: 05; Request string: sound convertor; Time search started: Wed Aug 27 16:49:39 BST 2008; Search term sound matches T2: sound; Score: 0.0; Search term convertor matches T1: windows; Score: 1.6540683337039817; Matched components: ID: 5 Score: 10| ID: 4 Score: 9| ID: 16 Score: 9| ID: 29 Score: 7| ID: 24 Score: 7| ID: 8 Score: 6| ID: 43 Score: 5| ID: 7 Score: 5| ID: 1 Score: 5| ID: 34 Score: 5| ID: 33 Score: 5| ID: 31 Score: 4| ID: 39 Score: 4| ID: 10 Score: 3| ID: 28 Score: 3| ID: 3 Score: 3| ID: 49 Score: 3| ID: 35 Score: 3| ID: 6 Score: 2| ID: 18 Score: 2| ID: 12 Score: 2| ID: 50 Score: 2| ID: 13 Score: 2| ID: 30 Score: 2| ID: 2 Score: 2| ID: 26 Score: 2| ID: 41 Score: 2| ID: 47 Score: 2| ID: 40 Score: 2| ID: 45 Score: 2| ID: 32 Score: 2| ID: 38 Score: 2| ID: 36 Score: 2| ID: 46 Score: 1| ID: 19 Score: 1| ID: 11 Score: 1| ID: 22 Score: 1| ID: 42 Score: 1| ID: 27 Score: 1| ID: 37 Score: 1| ID: 15 Score: 1| ID: 48 Score: 1| ID: 44 Score: 1| ID: 51 Score: 1| ID: 17 Score: 1| ID: 20 Score: 1| ; Time results returned: Wed Aug 27 16:50:27 BST 2008; User returned results at: Wed Aug 27 16:50:41 BST 2008; User section: 1,
Text Search:: Scenario NO: 01; Request string: file upload; Time search started: Fri Aug 29 13:42:37 BST 2008; ID: 4 Score: 1| ID: 6 Score: 1| ID: 8 Score: 1| ID: 11 Score: 1| ID: 16 Score: 1| ID: 39 Score: 1| ; Time results returned: Fri Aug 29 13:42:37 BST 2008; User returned results at: Fri Aug 29 13:42:44 BST 2008; User section: 6,
Text Search:: Scenario NO: 02; Request string: network traffic; Time search started: Fri Aug 29 13:42:59 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ; Time results returned: Fri Aug 29 13:42:59 BST 2008; User returned results at: Fri Aug 29 13:43:03 BST 2008; User section: 10,
Text Search:: Scenario NO: 03; Request string: remote desktopping; Time search started: Fri Aug 29 13:43:18 BST 2008; ID: 8 Score: 1| ID: 49 Score: 1| ; Time results returned: Fri Aug 29 13:43:18 BST 2008; User returned results at: Fri Aug 29 13:43:20 BST 2008; User section: 8,
Text Search:: Scenario NO: 04; Request string: mpeg player; Time search started: Fri Aug 29 13:43:33 BST 2008; ID: 3 Score: 1| ID: 11 Score: 1| ID: 27 Score: 1| ID: 38 Score: 1| ID: 51 Score: 1| ; Time results returned: Fri Aug 29 13:43:33 BST 2008; User returned results at: Fri Aug 29 13:43:40 BST 2008; User section: 11,

Text Search:: Scenario NO: 05; Request string: mp3 maker; Time search started: Fri Aug 29 13:43:52 BST 2008; ID: 1 Score: 1| ID: 3 Score: 1| ; Time results returned: Fri Aug 29 13:43:52 BST 2008; User returned results at: Fri Aug 29 13:43:57 BST 2008; User section: 1,
Experiment run 15,16,18 Aug 2008 Alternative:
faulty query rerun (refer to 1) above)
f>>>Text Search:: Scenario NO: 01; Request string: download onto server; Time search started: Fri Aug 29 14:00:45 BST 2008; ID: 46 Score: 1| ID: 6 Score: 1| ID: 13 Score: 1| ID: 24 Score: 1| ID: 27 Score: 1| ID: 29 Score: 1| ID: 34 Score: 1| ID: 44 Score: 1| ; Time results returned: Fri Aug 29 14:00:45 BST 2008; User returned results at: Fri Aug 29 14:00:56 BST 2008; User section: 6, 24,
SemaSearch:: Scenario NO: 01; Request string: download onto server; Time search started: Fri Aug 29 14:18:16 BST 2008; Search term download matches T1: http; Score: 0.678429415418877; Search term onto matches T1: integration; Score: 2.846099843450776; Search term server matches T2: server; Score: 0.0; Matched components: ID: 38 Score: 12| ID: 23 Score: 11| ID: 45 Score: 8| ID: 21 Score: 7| ID: 29 Score: 6| ID: 36 Score: 5| ID: 34 Score: 5| ID: 24 Score: 5| ID: 48 Score: 4| ID: 41 Score: 4| ID: 28 Score: 4| ID: 51 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 5 Score: 3| ID: 49 Score: 3| ID: 40 Score: 3| ID: 4 Score: 3| ID: 44 Score: 3| ID: 13 Score: 3| ID: 50 Score: 2| ID: 14 Score: 2| ID: 15 Score: 2| ID: 18 Score: 2| ID: 35 Score: 2| ID: 47 Score: 2| ID: 37 Score: 2| ID: 16 Score: 1| ID: 42 Score: 1| ID: 22 Score: 1| ID: 26 Score: 1| ID: 9 Score: 1| ID: 1 Score: 1| ID: 10 Score: 1| ID: 12 Score: 1| ID: 31 Score: 1| ID: 3 Score: 1| ID: 30 Score: 1| ID: 33 Score: 1| ID: 43 Score: 1| ID: 27 Score: 1| ; Time results returned: Fri Aug 29 14:19:19 BST 2008; User returned results at: Fri Aug 29 14:20:09 BST 2008; User section: 24, 6,
SemaSearch:: Scenario NO: 02; Request string: netwrok traffic; Time search started: Fri Aug 29 14:20:26 BST 2008; Search term netwrok matches T1: windows; Score: 0.9999802199297162; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Matched components: ID: 5 Score: 10| ID: 16 Score: 9| ID: 29 Score: 7| ID: 24 Score: 7| ID: 8 Score: 6| ID: 4 Score: 6| ID: 43 Score: 5| ID: 7 Score: 5| ID: 1 Score: 5| ID: 34 Score: 5| ID: 33 Score: 5| ID: 31 Score: 4| ID: 39 Score: 4| ID: 10 Score: 3| ID: 28 Score: 3| ID: 13 Score: 3| ID: 3 Score: 3| ID: 49 Score: 3| ID: 35 Score: 3| ID: 6 Score: 2| ID: 18 Score: 2| ID: 12 Score: 2| ID: 50 Score: 2| ID: 30 Score: 2| ID: 2 Score: 2| ID: 26 Score: 2| ID: 41 Score: 2| ID: 47 Score: 2| ID: 40 Score: 2| ID: 45 Score: 2| ID: 32 Score: 2| ID: 38 Score: 2| ID: 36 Score: 2| ID: 46 Score: 1| ID: 19 Score: 1| ID: 11 Score: 1| ID: 22 Score: 1| ID: 42 Score: 1| ID: 27 Score: 1| ID: 37 Score: 1| ID: 15 Score: 1| ID: 48 Score: 1| ID: 44 Score: 1| ID: 51 Score: 1| ID: 17 Score: 1| ID: 20 Score: 1| ; Time results returned: Fri Aug 29 14:21:16 BST 2008; User returned results at: Fri Aug 29 14:22:28 BST 2008; User section: 10,
SemaSearch:: Scenario NO: 03; Request string: remote control computer; Time search started: Fri Aug 29 14:22:46 BST 2008; Search term remote matches T3: remote; Score: 0.0; Search term control matches T3: control; Score: 0.0; Search term computer matches T3: mp; Score: 0.5343298238530242; Matched components: ID: 8 Score: 4| ID: 35 Score: 3| ID: 36 Score: 2| ID: 3 Score: 2| ID: 49 Score: 1| ID: 1 Score: 1| ; Time results returned: Fri Aug 29 14:24:22 BST 2008; User returned results at: Fri Aug 29 14:24:36 BST 2008; User section: 8,
SemaSearch:: Scenario NO: 04; Request string: media applications; Time search started: Fri Aug 29 14:24:50 BST 2008; Search term media matches T3: media; Score: 0.0; Search term applications matches T3: applications; Score: 0.0; Matched components: ID: 29 Score: 7| ID: 11 Score: 2| ID: 37 Score: 1| ; Time results returned: Fri Aug 29 14:26:10 BST 2008; User returned results at: Fri Aug 29 14:28:37 BST 2008; User section: 11,
SemaSearch:: Scenario NO: 05; Request string: modify sound files; Time search started: Fri Aug 29 14:28:52 BST 2008; Search term modify matches T1: administration; Score: 2.734179157544367; Search term sound matches T2: sound; Score: 0.0; Search term files matches T3: files; Score: 0.0; Matched components: ID: 4 Score: 9| ID: 16 Score: 9| ID: 39 Score: 4| ID: 47 Score: 2| ID: 35 Score: 2| ID: 26 Score: 1| ID: 50 Score: 1| ID: 28 Score: 1| ID: 34 Score: 1| ID: 1 Score: 1| ID: 43 Score: 1| ID: 22 Score: 1| ID: 11 Score: 1| ; Time results returned: Fri Aug 29 14:30:05 BST 2008; User returned results at: Fri Aug 29

14:33:07 BST 2008; User section: 1,
Text Search:: Scenario NO: 01; Request string: ftp client; Time search started: Fri Aug 29 15:14:18 BST 2008; ID: 6 Score: 2| ID: 24 Score: 2| ID: 13 Score: 1| ID: 34 Score: 1| ID: 46 Score: 1| ; Time results returned: Fri Aug 29 15:14:18 BST 2008; User returned results at: Fri Aug 29 15:14:21 BST 2008; User section: 6, 24,
Text Search:: Scenario NO: 02; Request string: monitor network traffic; Time search started: Fri Aug 29 15:14:33 BST 2008; ID: 10 Score: 1| ID: 34 Score: 1| ID: 51 Score: 1| ; Time results returned: Fri Aug 29 15:14:33 BST 2008; User returned results at: Fri Aug 29 15:14:37 BST 2008; User section: 10,
Text Search:: Scenario NO: 03; Request string: controling computer; Time search started: Fri Aug 29 15:14:49 BST 2008; ID: 5 Score: 1| ; Time results returned: Fri Aug 29 15:14:49 BST 2008; User returned results at: Fri Aug 29 15:14:52 BST 2008; User section:
Text Search:: Scenario NO: 04; Request string: video playback and audio playback; Time search started: Fri Aug 29 15:15:18 BST 2008; ID: 4 Score: 3| ID: 2 Score: 2| ID: 11 Score: 2| ID: 27 Score: 2| ID: 39 Score: 2| ID: 1 Score: 2| ID: 5 Score: 1| ID: 6 Score: 1| ID: 10 Score: 1| ID: 12 Score: 1| ID: 14 Score: 1| ID: 15 Score: 1| ID: 16 Score: 1| ID: 18 Score: 1| ID: 21 Score: 1| ID: 22 Score: 1| ID: 24 Score: 1| ID: 28 Score: 1| ID: 32 Score: 1| ID: 33 Score: 1| ID: 34 Score: 1| ID: 36 Score: 1| ID: 37 Score: 1| ID: 38 Score: 1| ID: 40 Score: 1| ID: 41 Score: 1| ID: 42 Score: 1| ID: 43 Score: 1| ID: 44 Score: 1| ID: 46 Score: 1| ID: 47 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 51 Score: 1| ID: 3 Score: 1| ; Time results returned: Fri Aug 29 15:15:18 BST 2008; User returned results at: Fri Aug 29 15:15:27 BST 2008; User section:
Text Search:: Scenario NO: 05; Request string: mp3 conversastion; Time search started: Fri Aug 29 15:15:45 BST 2008; ID: 1 Score: 1| ID: 3 Score: 1| ; Time results returned: Fri Aug 29 15:15:45 BST 2008; User returned results at: Fri Aug 29 15:15:48 BST 2008; User section: 1,
SemaSearch:: Scenario NO: 01; Request string: ssh secure shell; Time search started: Fri Aug 29 15:55:45 BST 2008; Search term ssh matches T2: telnet; Score: 0.5567197041232899; Search term secure matches T1: connection; Score: 1.083837100096054; Search term shell matches T2: unix; Score: 0.6807078266247116; Matched components: ID: 1 Score: 5| ID: 31 Score: 4| ID: 10 Score: 3| ID: 24 Score: 2| ID: 8 Score: 1| ID: 13 Score: 1| ID: 5 Score: 1| ID: 19 Score: 1| ID: 49 Score: 1| ID: 32 Score: 1| ; Time results returned: Fri Aug 29 15:56:55 BST 2008; User returned results at: Fri Aug 29 15:57:31 BST 2008; User section:
SemaSearch:: Scenario NO: 02; Request string: monitor network traffic; Time search started: Fri Aug 29 15:57:38 BST 2008; Search term monitor matches T1: servers; Score: 0.6294424386323639; Search term network matches T3: network; Score: 0.0; Search term traffic matches T2: proxy; Score: 1.0051551763652107; Matched components: ID: 24 Score: 5| ID: 10 Score: 5| ID: 51 Score: 5| ID: 13 Score: 4| ID: 6 Score: 4| ID: 8 Score: 4| ID: 34 Score: 3| ID: 29 Score: 1| ID: 44 Score: 1| ID: 27 Score: 1| ID: 48 Score: 1| ID: 49 Score: 1| ID: 31 Score: 1| ID: 30 Score: 1| ID: 14 Score: 1| ID: 18 Score: 1| ID: 39 Score: 1| ID: 36 Score: 1| ; Time results returned: Fri Aug 29 15:58:51 BST 2008; User returned results at: Fri Aug 29 15:59:19 BST 2008; User section: 10,
SemaSearch:: Scenario NO: 03; Request string: remotly control computer; Time search started: Fri Aug 29 15:59:48 BST 2008; Search term remotly matches T2: text; Score: 1.3974317037787685; Search term control matches T3: control; Score: 0.0; Search term computer matches T3: mp; Score: 0.5343298238530242; Matched components: ID: 35 Score: 3| ID: 31 Score: 2| ID: 48 Score: 2| ID: 36 Score: 2| ID: 3 Score: 2| ID: 8 Score: 1| ID: 46 Score: 1| ID: 38 Score: 1| ID: 1 Score: 1| ; Time results returned: Fri Aug 29 16:01:33 BST 2008; User returned results at: Fri Aug 29 16:01:50 BST 2008; User section:
SemaSearch:: Scenario NO: 04; Request string: media player to plsy video/audio; Time search started: Fri Aug 29 16:02:00 BST 2008; Search term media matches T3: media; Score: 0.0; Search term player matches T3: player; Score: 0.0; Search term to matches T3: io; Score: 0.05525913810672244; Search term plsy matches T1: lightweight; Score: 0.9998756962920724; Search term video/audio matches T3: system; Score: 1.4773071742090615; Matched components: ID: 46 Score: 7| ID: 30 Score: 6| ID: 21 Score: 6| ID: 43 Score: 5| ID: 51 Score: 5| ID: 11 Score: 4| ID: 22 Score: 3| ID: 25 Score: 3| ID: 36 Score: 2| ID: 7 Score: 1| ID: 33 Score: 1| ID: 16 Score: 1| ID: 48 Score: 1| ID: 35 Score: 1| ID: 47 Score: 1| ID: 37 Score: 1| ID: 20 Score: 1| ; Time results returned: Fri Aug 29 16:05:33

BST 2008; User returned results at: Fri Aug 29 16:06:28 BST 2008; User section: 11,
SemaSearch:: Scenario NO: 05; Request string: sound file modifier and converter; Time search started: Fri Aug 29 16:06:43 BST 2008; Search term sound matches T2: sound; Score: 0.0; Search term file matches T3: file; Score: 0.0; Search term modifier matches T1: site; Score: 0.523090538333863; Search term and matches T3: io; Score: 0.05070387634032431; Search term converter matches T2: formats; Score: 1.5196788319220156; Matched components: ID: 30 Score: 12| ID: 39 Score: 8| ID: 37 Score: 8| ID: 35 Score: 8| ID: 49 Score: 7| ID: 21 Score: 6| ID: 46 Score: 6| ID: 44 Score: 6| ID: 38 Score: 6| ID: 1 Score: 6| ID: 4 Score: 5| ID: 51 Score: 5| ID: 43 Score: 5| ID: 23 Score: 4| ID: 34 Score: 4| ID: 42 Score: 4| ID: 15 Score: 4| ID: 24 Score: 4| ID: 11 Score: 4| ID: 50 Score: 3| ID: 25 Score: 3| ID: 22 Score: 3| ID: 14 Score: 3| ID: 48 Score: 3| ID: 3 Score: 3| ID: 47 Score: 3| ID: 32 Score: 2| ID: 45 Score: 2| ID: 12 Score: 2| ID: 41 Score: 2| ID: 6 Score: 2| ID: 16 Score: 2| ID: 17 Score: 2| ID: 8 Score: 2| ID: 9 Score: 2| ID: 40 Score: 2| ID: 10 Score: 2| ID: 29 Score: 2| ID: 7 Score: 2| ID: 18 Score: 2| ID: 20 Score: 1| ID: 26 Score: 1| ID: 5 Score: 1| ID: 33 Score: 1| ID: 31 Score: 1| ID: 36 Score: 1| ID: 27 Score: 1| ; Time results returned: Fri Aug 29 16:09:43 BST 2008; User returned results at: Fri Aug 29 16:09:56 BST 2008; User section: 1,

A3: SourceForge.net case study log analyses

#	System	Scenario	Search text	Matched Ids	Scenario Ids	# Scen Ids	Res count	Irrelevant	Relevant	Missed	Precision	Recall	F-Score
1	Text	1	download file server	6,46,4,8,11,16,39,13,24,27,29,34,44	6,8,24	3	13	10	3	0	0.230769231	1	0.375
2	Sema CS	1	download file server	38,23,21,45,29,34,24,48,41,39,6,8	6,8,24	3	12	9	3	0	0.25	1	0.4
3	Sema CS	1	ftp upload software	24,8,29,6,27,43,47	6,8,24	3	7	4	3	0	0.428571429	1	0.6
4	Text	1	ftp upload software	24,6,8,27,29,4,47	6,8,24	3	7	4	3	0	0.428571429	1	0.6
5	Sema CS	2	network traffic monitoring tools	10,51,24,13,6,8	10	1	6	5	1	0	0.166666667	1	0.285714286
6	Text	2	network traffic monitoring tools	10,34,51, 15,35	10	1	5	4	1	0	0.2	1	0.333333333
7	Text	2	network traffic monitor	10,34,51	10	1	3	2	1	0	0.333333333	1	0.5
8	Sema CS	2	network traffic monitor	10,51,24,13,6,8	10	1	6	5	1	0	0.166666667	1	0.285714286
9	Sema CS	3	software download free remote PC control	29,8,38,23,24,21,45,43,36,41,18,35	8	1	12	11	1	0	0.083333333	1	0.153846154
10	Text	3	software download free remote PC control	8,29,24,27,43,47,46,22,49,14,18,48,4,19,35,36	8	1	16	15	1	0	0.0625	1	0.117647059
11	Text	3	remote application to control computer	49,8,2,7,29,30,37,42,4,5,35,3,1,10,13,14,15,16,17,18,22,23,24,26,28,31,33,34,38,39,40,41,43,44,45,46,47,48,50,19,36	8	1	41	40	1	0	0.024390244	1	0.047619048
12	Sema CS	3	remote application to control computer	30,3,7,8,49	8	1	5	4	1	0	0.2	1	0.333333333

13	Sema CS	4	video audio player	3,1,11,27	1,3,4,11,27	5	4	0	4	1	1	0.8	0.888888889
14	Text	4	video audio player	11,27,2,39,1,3,38,51	1,3,4,11,27	5	8	4	4	1	0.5	0.8	0.615384615
15	Text	4	mpeg4	Null	1,3,4,11,27	5	0	0	0	5	0	0	0
16	Sema CS	4	mpeg4	7,3,49	1,3,4,11,27	5	3	2	1	4	0.333333333	0.2	0.25
17	Sema CS	5	sound conversion	1,4,11	1	1	3	2	1	0	0.333333333	1	0.5
18	Text	5	sound conversion	4	1	1	1	1	0	1	0	0	0
19	Text	5	sound file converter	4,6,8,11,16,39	1	1	6	6	0	1	0	0	0
20	Sema CS	5	sound file converter	1,39	1	1	2	1	1	0	0.5	1	0.666666667
21	Sema CS	1	software to upload files	16,8,24,29,4,27,43,30,47,39	6,8,24	3	10	8	2	1	0.2	0.666666667	0.307692308
22	Sema CS	2	network traffic monitor tool	24,49,10,51,13,6,8	10	1	7	6	1	0	0.142857143	1	0.25
23	Sema CS	3	remote control	8,35,36	8	1	3	2	1	0	0.333333333	1	0.5
24	Sema CS	4	audio	3,1	1,3,4,11,27	5	2	0	2	3	1	0.4	0.571428571
25	Sema CS	5	sound file conversion	1,39	1	1	2	1	1	0	0.5	1	0.666666667
26	Text	1	software to upload files	8,24,29,43,47,4,16,39,27,1,2,5,7,10,13,14,15,17,18,22,23,26,28,30,31,33,34,35,37,38,40,41,42,44,45,46,48,49,50,11	6,8,24	3	40	38	2	1	0.05	0.666666667	0.093023256
27	Text	2	network traffic monitor tool	10,34,51,15,17,24,29,35,49	10	1	9	8	1	0	0.111111111	1	0.2
28	Text	3	remote control	8,49,4,19,35,36	8	1	6	5	1	0	0.166666667	1	0.285714286
29	Text	4	audio	1,3,11,27	1,3,4,11,27	5	4	0	4	1	1	0.8	0.888888889

30	Text	5	sound file conversion	4,6,8,11,16,39	1	1	6	6	0	1	0	0	0
31	Text	1	application dowload upload server	29,2,3,7,30,37,42,49,6,13,24,27,34,44	6,8,24	3	14	12	2	1	0.142857143	0.666666667	0.235294118
32	Text	2	monitor network traffic	10,34,51	10	1	3	2	1	0	0.333333333	1	0.5
33	Text	3	application for remotly control computer	7,29,42,49,4,5,35,36,2,3,30,37,1,6,10,11,12,15,16,18,21,23,26,27,31,34,41,44,45,47,51,19	8	1	33	33	0	1	0	0	0
34	Text	4	media player	11,27,38,51	1,3,4,11,27	5	4	2	2	3	0.5	0.4	0.444444444
35	Text	5	sound convertor	4	1	1	1	1	0	1	0	0	0
36	Sema CS	1	application dowload upload server	27,49,46,6,3,7,42,35,51,48,8,24,23,4,38,10,13	6,8,24	3	17	14	3	0	0.176470588	1	0.3
37	Sema CS	2	monitor network traffic	24,10,51,13,6,8	10	1	6	5	1	0	0.166666667	1	0.285714286
38	Sema CS	3	application for remotly control computer	16,4,3,7	8	1	4	4	0	1	0	0	0
39	Sema CS	4	media player	4	1,3,4,11,27	5	1	0	1	4	1	0.2	0.333333333
40	Sema CS	5	sound convertor	5,4,16,29,24,8,43,7,1,34,33,31,39	1	1	13	12	1	0	0.076923077	1	0.142857143
41	Sema CS	1	file upload	27, 39	6,8,24	3	2	2	0	3	0	0	0
42	Sema CS	2	network traffic	51,10	10	1	2	1	1	0	0.5	1	0.666666667
43	Sema CS	3	remote desktopping	8	8	1	1	0	1	0	1	1	1
44	Sema CS	4	mpeg player	3	1,3,4,11,27	5	1	0	1	4	1	0.2	0.333333333

45	Sema CS	5	mp3 maker	3,1,24,6,8	1	1	5	4	1	0	0.2	1	0.333333333
46	Text	1	file upload	4,6,8,11,16,39	6,8,24	3	6	4	2	1	0.333333333	0.666666667	0.444444444
47	Text	2	network traffic	10,34,51	10	1	3	2	1	0	0.333333333	1	0.5
48	Text	3	remote desktopping	8,49	8	1	2	1	1	0	0.5	1	0.666666667
49	Text	4	mpeg player	3,11,27,38,51	1,3,4,11,27	5	5	2	3	2	0.6	0.6	0.6
50	Text	5	mp3 maker	1,3	1	1	2	1	1	0	0.5	1	0.666666667
51	Text	1	download onto server	46,6,13,24,27,29,34,44	6,8,24	3	8	6	2	1	0.25	0.666666667	0.363636364
52	Text	2	netwrok traffic	null	10	1	0	0	0	1	0	0	0
53	Text	3	remote control computer	8,49,4,19,35,36,5	8	1	7	6	1	0	0.142857143	1	0.25
54	Text	4	media applications	11,29,37	1,3,4,11,27	5	3	2	1	4	0.333333333	0.2	0.25
55	Text	5	modify sound files	39,4,11,16	1	1	4	4	0	1	0	0	0
56	Sema CS	1	download onto server	38,23,45,21,29,36,34,24,48,41,28,51,6,8	6,8,24	3	14	11	3	0	0.214285714	1	0.352941176
57	Sema CS	2	netwrok traffic	5,16,29,24,8,4,43,7,1,34,33,31,39	10	1	13	13	0	1	0	0	0
58	Sema CS	3	remote control computer	8	8	1	1	0	1	0	1	1	1
59	Sema CS	4	media applications	29	1,3,4,11,27	5	1	1	0	5	0	0	0
60	Sema CS	5	modify sound files	4,39	1	1	2	2	0	1	0	0	0
61	Sema CS	1	ftp client	24,6,34	6,8,24	3	3	1	2	1	0.666666667	0.666666667	0.666666667
62	Sema CS	2	monitor network traffic	24,10,51,13,6,8	10	1	6	5	1	0	0.166666667	1	0.285714286

63	Sema CS	3	controling computer	6, 10	8	1	2	2	0	1	0	0	0
64	Sema CS	4	video playback and audio playback	3,1,4,11,30,27	1,3,4,11,27	5	6	1	5	0	0.833333333	1	0.909090909
65	Sema CS	5	mp3 conversastion	3,1	1	1	2	1	1	0	0.5	1	0.666666667
66	Text	1	ftp client	6,24,13,34,46	6,8,24	3	5	3	2	1	0.4	0.666666667	0.5
67	Text	2	monitor network traffic	10,34,51	10	1	3	2	1	0	0.333333333	1	0.5
68	Text	3	controling computer	5	8	1	1	1	0	1	0	0	0
69	Text	4	video playback and audio playback	4,2,11,27,39,1,5,6,10,12,14,15,16,18,21,22,24,28,32,33,34,36,37,38,40,41,42,43,44,46,47,48,49,51,3	1,3,4,11,27	5	35	30	5	0	0.142857143	1	0.25
70	Text	5	mp3 conversastion	1,3	1	1	2	1	1	0	0.5	1	0.666666667
71	Text	1	ssh secure shell	21	6,8,24	3	1	1	0	3	0	0	0
72	Text	2	monitor network traffic	10,34,51	10	1	3	2	1	0	0.333333333	1	0.5
73	Text	3	remotly control computer	4,19,35,36,5	8	1	5	5	0	1	0	0	0
74	Text	4	media player to plsy video/audio	11,38,27,51,1,2,4,5,7,8,10,13,14,15,16,17,18,22,23,24,26,28,29,30,31,33,34,35,37,39, 40,41,42,43,44,45,46,47,48,49, 50	1,3,4,11,27	5	41	37	4	1	0.097560976	0.8	0.173913043
75	Text	5	sound file modifier and converter	4,6,16,39,8,11,1,2,5,10,12,14,15,18,21,22,24,28,32,33,34,36,37,38,40,41,42,43,44,46,47,48,49,51	1	1	34	33	1	0	0.029411765	1	0.057142857
76	Sema CS	1	ssh secure shell	1,31	6,8,24	3	2	2	0	3	0	0	0
77	Sema CS	2	monitor network traffic	24,10,51,13,6,8	10	1	6	5	1	0	0.166666667	1	0.285714286
78	Sema CS	3	remotly control computer	35	8	1	1	1	0	1	0	0	0

79	Sema CS	4	media player to plsy video/audio	46,30,21,43,51,11	1,3,4,11,27	5	5	5	0	5	0	0	0
80	Sema CS	5	sound file modifier and converter	30,39,37,35,49,21,46,44,38,1,4,51,43,23,34,42,15,24,11	1	1	19	18	1	0	0.052631579	1	0.1

A3 A: SemaCS search interpolated average P/R and MRR

SemaCS search: scenario 1 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
2	38,23,21,45,29,34,24,48,41,39,6,8	12	3	6,8,24	3	7; 11; 12	1.000000	0.250000	0.000000
3	24,8,29,6,27,43,47	7	3	6,8,24	3	1; 2; 4	1.000000	0.428571	1.000000
21	16,8,24,29,4,27,43,30,47,39	10	3	6,8,24	2	2; 3	0.666667	0.200000	0.500000
36	27,49,46,6,3,7,42,35,51,48,8,24,23,4,38,10,13	17	3	6,8,24	3	4; 11; 12	1.000000	0.176471	0.250000
41	27, 39	2	3	6,8,24	0	0	0.000000	0.000000	0.000000
56	38,23,45,21,29,36,34,24,48,41,28,51,6,8	14	3	6,8,24	3	8; 13; 14	1.000000	0.214286	0.000000
61	24,6,34	3	3	6,8,24	2	1; 2	0.666667	0.666667	0.500000
76	1,31	2	3	6,8,24	0	0	0.000000	0.000000	0.000000
					Scenario 1 Average:		0.666667	0.241999	0.281250

	SemaCS search: scenario 1 (8 queries) Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 2 Prec				**0.14**			**0.18**				**0.25**
Q 3 Prec				**1.00**			**1.00**				**0.75**
Q 21 Prec				**0.50**			**0.67**				**0.00**
Q 36 Prec				**0.25**			**0.18**				**0.25**
Q 41 Prec				**0.00**			**0.00**				**0.00**
Q 56 Prec				**0.13**			**0.15**				**0.21**
Q 61 Prec				**1.00**			**1.00**				**0.00**
Q 76 Prec				**0.00**			**0.00**				**0.00**

	SemaCS search: scenario 1 (8 queries) Interpolated Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 2 Prec	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Q 3 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	0.75	0.75	0.75
Q 21 Prec	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.00	0.00	0.00	0.00
Q 36 Prec	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Q 41 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 56 Prec	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
Q 61 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Q 76 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.18	0.18	0.18	0.18

SemaCS search: scenario 2 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
5	10,51,24,13,6,8	6	1	10	1	1	1.000000	0.166667	1.000000
8	10,51,24,13,6,8	6	1	10	1	1	1.000000	0.166667	1.000000
22	24,49,10,51,13,6,8	7	1	10	1	3	1.000000	0.142857	0.333333
37	24,10,51,13,6,8	6	1	10	1	2	1.000000	0.166667	0.500000
42	51,10	2	1	10	1	2	1.000000	0.500000	0.500000
57	5,16,29,24,8,4,43,7,1,34,33,31,39	13	1	10	0	0	0.000000	0.000000	0.000000
62	24,10,51,13,6,8	6	1	10	1	2	1.000000	0.166667	0.500000
77	24,10,51,13,6,8	6	1	10	1	2	1.000000	0.166667	0.500000
					Scenario 2 Average:		**0.875000**	**0.184524**	**0.541667**

	SemaCS search: scenario 2 (8 queries) Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 5 Prec											**1.00**
Q 8 Prec											**1.00**
Q 22 Prec											**0.33**
Q 37 Prec											**0.50**
Q 42 Prec											**0.50**
Q 57 Prec											**0.00**
Q 62 Prec											**0.50**
Q 77 Prec											**0.50**

	SemaCS search: scenario 2 (8 queries) Interpolated Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 5 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 8 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 22 Prec	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Q 37 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 42 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 57 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 62 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 77 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Average	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54

SemaCS search: scenario 3 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
9	29,8,38,23,24,21,45,43,36,41,18,35	12	1	8	1	2	1.000000	0.083333	0.500000
12	30,3,7,8,49	5	1	8	1	4	1.000000	0.200000	0.250000
23	8,35,36	3	1	8	1	1	1.000000	0.333333	1.000000
38	16,4,3,7	4	1	8	0	0	0.000000	0.000000	0.000000
43	8	1	1	8	1	1	1.000000	1.000000	1.000000
58	8	1	1	8	1	1	1.000000	1.000000	1.000000
63	6, 10	2	1	8	0	0	0.000000	0.000000	0.000000
78	35	1	1	8	0	0	0.000000	0.000000	0.000000
					Scenario 3 Average:		**0.625000**	**0.327083**	**0.468750**

	SemaCS search: scenario 3 (8 queries) Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 9 Prec											**0.50**
Q 12 Prec											**0.25**
Q 23 Prec											**1.00**
Q 38 Prec											**0.00**
Q 43 Prec											**1.00**
Q 58 Prec											**1.00**
Q 63 Prec											**0.00**
Q 78 Prec											**0.00**

	SemaCS search: scenario 3 (8 queries) Interpolated Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 9 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 12 Prec	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Q 23 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 38 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 43 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 58 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 63 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 78 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**

SemaCS search: scenario 4 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
13	3,1,11,27	4	5	1,3,4,11,27	4	1,2,3,4	0.800000	1.000000	1.000000
16	7,3,49	3	5	1,3,4,11,27	1	2	0.200000	0.333333	0.500000
24	3,1	2	5	1,3,4,11,27	2	1,2	0.400000	1.000000	1.000000
39	4	1	5	1,3,4,11,27	1	1	0.200000	1.000000	1.000000
44	3	1	5	1,3,4,11,27	1	1	0.200000	1.000000	1.000000
59	29	1	5	1,3,4,11,27	0	0	0.000000	0.000000	0.000000
64	3,1,4,11,30,27	6	5	1,3,4,11,27	5	1,2,3,4,6	1.000000	0.833333	1.000000
79	46,30,21,43,51,11	6	5	1,3,4,11,27	1	6	0.200000	0.200000	0.000000
					Scenario 4 Average:		**0.375000**	**0.670833**	**0.687500**

SemaCS search: scenario 4 (8 queries) Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 13 Prec			**1.00**		**1.00**		**1.00**		**1.00**		**0.00**
Q 16 Prec			**0.50**		**0.00**		**0.00**		**0.00**		**0.00**
Q 24 Prec			**1.00**		**1.00**		**0.00**		**0.00**		**0.00**
Q 39 Prec			**1.00**		**0.00**		**0.00**		**0.00**		**0.00**
Q 44 Prec			**1.00**		**0.00**		**0.00**		**0.00**		**0.00**
Q 59 Prec			**0.00**		**0.00**		**0.00**		**0.00**		**0.00**
Q 64 Prec			**1.00**		**1.00**		**1.00**		**0.83**		**0.00**
Q 79 Prec			**0.17**		**0.00**		**0.00**		**0.00**		**0.00**

SemaCS search: scenario 4 (8 queries) Interpolated Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 13 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
Q 16 Prec	0.50	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 24 Prec	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 39 Prec	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 44 Prec	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 59 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 64 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.83	0.00	0.00
Q 79 Prec	0.17	0.17	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	0.71	0.71	0.71	0.38	0.38	0.25	0.25	0.23	0.23	0.00	0.00

SemaCS search: scenario 5 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
17	1,4,11	3	1	1	1	1	1.000000	0.333333	1.000000
20	1,39	2	1	1	1	1	1.000000	0.500000	1.000000
25	1,39	2	1	1	1	1	1.000000	0.500000	1.000000
40	5,4,16,29,24,8,43,7,1,34,33,31,39	13	1	1	1	9	1.000000	0.076923	0.000000
45	3,1,24,6,8	5	1	1	1	2	1.000000	0.200000	0.500000
60	4,39	2	1	1	0	0	0.000000	0.000000	0.000000
65	3,1	2	1	1	1	2	1.000000	0.500000	0.500000
80	30,39,37,35,49,21,46,44,38,1,4,51,43,23,34,42,15,24,11	19	1	1	1	10	1.000000	0.052632	0.000000
					Scenario 5 Average:		**0.875000**	**0.270361**	**0.500000**

	SemaCS search: scenario 5 (8 queries) Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 17 Prec											**1.00**
Q 20 Prec											**1.00**
Q 25 Prec											**1.00**
Q 40 Prec											**0.11**
Q 45 Prec											**0.50**
Q 60 Prec											**0.00**
Q 65 Prec											**0.50**
Q 80 Prec											**0.10**

	SemaCS search: scenario 5 (8 queries) Interpolated Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 17 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 20 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 25 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 40 Prec	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
Q 45 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 60 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 65 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 80 Prec	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Average	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**

A3 B: Text-based search interpolated average P/R and MRR

Text-based search: scenario 1 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
1	6,46,4,8,11,16,39,13,24,27,29,34,44	13	3	6,8,24	3	1, 4, 9	1.000000	0.230769	1.000000
4	24,6,8,27,29,4,47	7	3	6,8,24	3	1, 2, 3	1.000000	0.428571	1.000000
26	8,24,29,43,47,4,16,39,27,1,2,5,7,10,13,14,15,17,18,22,23,26, 28,30,31,33,34,35,37,38,40,41,42,44,45,46,48,49,50,11	40	3	6,8,24	2	1, 2	0.666667	0.050000	1.000000
31	29,2,3,7,30,37,42,49,6,13,24,27,34,44	14	3	6,8,24	2	9, 11	0.666667	0.142857	0.000000
46	4,6,8,11,16,39	6	3	6,8,24	2	2, 3	0.666667	0.333333	0.500000
51	46,6,13,24,27,29,34,44	8	3	6,8,24	2	2, 4	0.666667	0.250000	0.500000
66	6,24,13,34,46	5	3	6,8,24	2	1, 2	0.666667	0.400000	1.000000
71	21	1	3	6,8,24	0	0	0.000000	0.000000	0.000000
					Scenario 1 Average:		**0.666667**	**0.229441**	**0.625000**

	Text-based search: scenario 1 (8 queries) Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec				**1.00**				**0.50**			**0.33**
Q 4 Prec				**1.00**				**1.00**			**1.00**
Q 26 Prec				**1.00**				**1.00**			**0.00**
Q 31 Prec				**0.11**				**0.18**			**0.00**
Q 46 Prec				**0.50**				**0.67**			**0.00**
Q 51 Prec				**0.50**				**0.50**			**0.00**
Q 66 Prec				**1.00**				**1.00**			**0.00**
Q 71 Prec				**0.00**				**0.00**			**0.00**

	Text-based search: scenario 1 (8 queries) Interpolated Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec	1.00	1.00	1.00	1.00	0.50	0.50	0.50	0.50	0.33	0.33	0.33
Q 4 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 26 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
Q 31 Prec	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.00	0.00	0.00
Q 46 Prec	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.00	0.00	0.00
Q 51 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.00	0.00	0.00
Q 66 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
Q 71 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	**0.67**	**0.67**	**0.67**	**0.67**	**0.61**	**0.61**	**0.61**	**0.61**	**0.17**	**0.17**	**0.17**

Text-based search: scenario 2 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
6	10,34,51,15,35	5	1	10	1	1	1.000000	0.200000	1.000000
7	10,34,51	3	1	10	1	1	1.000000	0.333333	1.000000
27	10,34,51,15,17,24,29,35,49	9	1	10	1	1	1.000000	0.111111	1.000000
32	10,34,51	3	1	10	1	1	1.000000	0.333333	1.000000
47	10,34,51	3	1	10	1	1	1.000000	0.333333	1.000000
52	null	0	1	10	0	0	0.000000	0.000000	0.000000
67	10,34,51	3	1	10	1	1	1.000000	0.333333	1.000000
72	10,34,51	3	1	10	1	1	1.000000	0.333333	1.000000
					Scenario 2 Average:		**0.875000**	**0.247222**	**0.875000**

	Text-based search: scenario 2 (8 queries) Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 6 Prec											**1.00**
Q 7 Prec											**1.00**
Q 27 Prec											**1.00**
Q 32 Prec											**1.00**
Q 47 Prec											**1.00**
Q 52 Prec											**0.00**
Q 67 Prec											**1.00**
Q 72 Prec											**1.00**

	Text-based search: scenario 2 (8 queries) Interpolated Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 6 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 7 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 27 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 32 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 47 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 52 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 67 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 72 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**

Text-based search: scenario 3 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
10	8,29,24,27,43,47,46,22,49,14,18,48,4,19,35,36	16	1	8	1	1	1.000000	0.062500	1.000000
11	49,8,2,7,29,30,37,42,4,5,35,3,1,10,13,14,15,16,17,18,22,23,24,26, 28,31,33,34,38,39,40,41,43,44,45,46,47,48,50, 19,36	41	1	8	1	2	1.000000	0.024390	0.500000
28	8,49,4,19,35,36	6	1	8	1	1	1.000000	0.166667	1.000000
33	7,29,42,49,4,5,35,36,2,3,30,37,1,6,10,11,12,15,16,18,21,23,26,27, 31,34,41,44,45,47,51,19	33	1	8	0	0	0.000000	0.000000	0.000000
48	8,49	2	1	8	1	1	1.000000	0.500000	1.000000
53	8,49,4,19,35,36,5	7	1	8	1	1	1.000000	0.142857	1.000000
68	5	1	1	8	0	0	0.000000	0.000000	0.000000
73	4,19,35,36,5	5	1	8	0	0	0.000000	0.000000	0.000000
					Scenario 3 Average:		**0.625000**	**0.112052**	**0.562500**

	Text-based search: scenario 3 (8 queries) Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 10 Prec											**1.00**
Q 11 Prec											**0.50**
Q 28 Prec											**1.00**
Q 33 Prec											**0.00**
Q 48 Prec											**1.00**
Q 53 Prec											**1.00**
Q 68 Prec											**0.00**
Q 73 Prec											**0.00**

	Text-based search: scenario 3 (8 queries) Interpolated Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 10 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 11 Prec	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Q 28 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 33 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 48 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 53 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 68 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 73 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**	**0.56**

Text-based search: scenario 4 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
14	11,27,2,39,1,3,38,51	8	5	1,3,4,11,27	4	1,2,5,6	0.800000	0.500000	1.000000
15	Null	0	5	1,3,4,11,27	0	0	0.000000	0.000000	0.000000
29	1,3,11,27	4	5	1,3,4,11,27	4	1,2,3,4	0.800000	1.000000	1.000000
34	11,27,38,51	4	5	1,3,4,11,27	2	1,2	0.400000	0.500000	1.000000
49	3,11,27,38,51	5	5	1,3,4,11,27	3	1,2,3	0.600000	0.600000	1.000000
54	11,29,37	3	5	1,3,4,11,27	1	1	0.200000	0.333333	1.000000
69	4,2,11,27,39,1,5,6,10,12,14,15,16,18,21,22,24,28,32,33,34,36, 37,38,40,41,42,43,44,46,47,48,49,51,3	35	5	1,3,4,11,27	5	1,3,4,6,35	1.000000	0.142857	1.000000
74	11,38,27,51,1,2,4,5,7,8,10,13,14,15,16,17,18,22,23,24,26,28,29, 30,31,33,34,35,37,39,40,41,42,43,44,45,46,47,48, 49,50	41	5	1,3,4,11,27	4	1,3,5,7	0.800000	0.097561	1.000000
					Scenario 4 Average:		**0.575000**	**0.396719**	**0.875000**

	Text-based search: scenario 4 (8 queries) Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 14 Prec			**1.00**		**1.00**		**0.60**		**0.67**		**0.00**
Q 15 Prec			**0.00**		**0.00**		**0.00**		**0.00**		**0.00**
Q 29 Prec			**1.00**		**1.00**		**1.00**		**1.00**		**0.00**
Q 34 Prec			**1.00**		**1.00**		**0.00**		**0.00**		**0.00**
Q 49 Prec			**1.00**		**1.00**		**1.00**		**0.00**		**0.00**
Q 54 Prec			**1.00**		**0.00**		**0.00**		**0.00**		**0.00**
Q 69 Prec			**1.00**		**0.67**		**0.75**		**0.67**		**0.14**
Q 74 Prec			**1.00**		**0.67**		**0.60**		**0.57**		**0.00**

	Text-based search: scenario 4 (8 queries) Interpolated Precision										
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 14 Prec	1.00	1.00	1.00	1.00	1.00	0.67	0.67	0.67	0.67	0.00	0.00
Q 15 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 29 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
Q 34 Prec	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 49 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Q 54 Prec	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 69 Prec	1.00	1.00	1.00	0.75	0.75	0.75	0.75	0.67	0.67	0.14	0.14
Q 74 Prec	1.00	1.00	1.00	0.67	0.67	0.60	0.60	0.57	0.57	0.00	0.00
Average	**0.88**	**0.88**	**0.88**	**0.68**	**0.68**	**0.50**	**0.50**	**0.36**	**0.36**	**0.02**	**0.02**

Text-based search: scenario 5 (8 queries) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
18	4	1	1	1	0	0	0.000000	0.000000	0.000000
19	4,6,8,11,16,39	6	1	1	0	0	0.000000	0.000000	0.000000
30	4,6,8,11,16,39	6	1	1	0	0	0.000000	0.000000	0.000000
35	4	1	1	1	0	0	0.000000	0.000000	0.000000
50	1,3	2	1	1	1	1	1.000000	0.500000	1.000000
55	39,4,11,16	4	1	1	0	0	0.000000	0.000000	0.000000
70	1,3	2	1	1	1	1	1.000000	0.500000	1.000000
75	4,6,16,39,8,11,1,2,5,10,12,14,15,18,21,22,24,28,32,33,34, 36,37,38,40,41,42,43,44,46,47,48,49,51	34	1	1	1	7	1.000000	0.029412	0.000000
					Scenario 5 Average:		**0.375000**	**0.128676**	**0.250000**

	Text-based search: scenario 5 (8 queries) Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 18 Prec											**0.00**
Q 19 Prec											**0.00**
Q 30 Prec											**0.00**
Q 35 Prec											**0.00**
Q 50 Prec											**1.00**
Q 55 Prec											**0.00**
Q 70 Prec											**1.00**
Q 75 Prec											**0.14**

	Text-based search: scenario 5 (8 queries) Interpolated Precision										
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Q 18 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 19 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 30 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 35 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 50 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 55 Prec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q 70 Prec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Q 75 Prec	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Average	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**

A4: SourceForge.net study participant information

What is SemaCS?

SemaCS is an innovative search engine currently referencing a small set of SourceForge.net software component descriptions.

Why do you need my help?

Your help is needed to evaluate SemaCS, its accuracy and efficiency, as well as to identify any further possible improvements.

What is required of me?

You will be provided with a search scenario. You are to use your own interpretation of the scenario to search for and identify software component ID (s) – and that's it!

How long would it take?

All scenarios are quite simple (really!) – Only about 5 minutes of your time would be required!

Why would I want to donate 5 minutes of my time?

Many reasons really: you would be helping to drive science, you would get a chance to try SemaCS (and you would be one of the first people to do so, you get a nice warm feeling knowing that you have made a difference! And you get to help an unfortunate PhD student looking for test subjects ☺

What data about me are you going to store?

In simple terms – no personal data of any kind is to be collected and/or stored! We simply log your query, time taken and component ID (s) that you feel are a match to your scenario – that is all!

I am really interested – can I have the results? Ask questions?

Of course! Once study data has been analysed all results will be made available! And if you have any questions – do feel free to ask or to get in touch via sjachym@wmin.ac.uk (after you have finished with your scenario though – we do not want you to be biased in any way!)

Appendix B: University of Westminster SRS case study

University of Westminster SRS study Total Years 1 and 2 interpolated average Precision Recall											
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
SemaCS no personalisation	0.5114	0.5114	0.5114	0.4876	0.4398	0.4398	0.2869	0.2632	0.2132	0.2137	0.2140
SemaCS with personalisation	0.5867	0.5867	0.5824	0.5744	0.5656	0.5191	0.3984	0.3601	0.3373	0.3019	0.2806
UoW SRS search	0.4638	0.4638	0.4638	0.4161	0.3468	0.3468	0.2319	0.1884	0.1429	0.1429	0.1429

University of Westminster SRS Total Years 1 and 2 study average				
	Recall	**Precision**	**MRR**	**F-score**
SemaCS no personalisation	0.496670	0.367139	0.503623	0.422193
SemaCS with personalisation	0.665657	0.422106	0.570393	0.516616
UoW SRS search	0.285533	0.463768	0.463768	0.353452

Appendix B1: University of Westminster SRS Year 1 data

mid	id	level	name	desc
1	3ECE508	5	PROFESSIONAL ENGINEERING PRACTICE	To give an appreciation of the contexts in which engineering knowledge and skills are applied in the workplace and prepare the student for professional career development and enhancement. To develop familiarity with professional equipment to a level where basic operation can be achieved.
2	3ECE514	5	EMBEDDED MICROPROCESSOR SYSTEM PROJECT	To supply the techniques required to design and commission an electrmechanical system controlled by an embedded microprocessor. To provide the challenge and problems of taking a design from conception through to having a working prototype. To give experience of working together in a team to achieve a common objective. To allow the use of development tools such as in-circuit emulators.
3	3ECE515	5	ANALOG ELECTRONIC DESIGN PROJECT	To give an understanding of various building blocks of analog electronic circuits together with examples of their application and implementation. To provide, within the context of a real application, experiencing of designing, simulating, prototyping and laboratory testing of real circuits. To provide further experience of working with others in a team to achieve a common objective.
4	3ECE516	5	DATA TRANSMISSION PROJECT	To apply the programming skills learned in the first year and new communication concepts introduced in this module to the design and construction of elements interfaced to a microprocessor. To give an understanding of the importance of subsystems division and specification.
5	3ECE517	5	DIGITAL MICROELECTRONIC DESIGN PROJECT	To introduce custom VLSI design on silicon in the MOS medium (primarily CMOS); to provide the student with a full design cycle experience from initial specification to simulation, physical layout and eventually the testing of fabricated devices; to create awareness of the factors that influence choice of technology, design style, fabrication route and CAD tools are addressed.
6	3ECE521	5	SIGNAL AND SYSTEM ANALYSIS	
7	3ECE522	5	COMMUNICATIONS AND CIRCUITS	
8	3ECE523	5	ENGINEERING SOFTWARE	Extends students' knowledge of high-level programming to implementation of larger, partitioned systems, exploring real-time requirements and relationship with operating systems.
9	3IIS503	5	BUSINESS ORGANISATION (SRI LANKA)	

10	3IIS509	5	PROJECT MANAGEMENT (SRI LANKA)	
11	3IIS551	5	DISTRIBUTED INFORMATION SYSTEMS (SRI LANKA)	
12	3IIS552	5	REQUIREMENTS ANALYSIS (SRI LANKA)	
13	3IIS553	5	DATABASE MANAGEMENT SYSTEMS (SRI LANKA)	
14	3IIS555	5	RAPID APPLICATION DEVELOPMENT (SRI LANKA)	
15	3IIS557	5	INTRODUCTION TO PROGRAMMING (SRI LANKA)	
16	3IIS558	5	BUSINESS ORGANISATION AND COMMUNICATION (SRI LANKA)	
17	3IIS560	5	MANAGEMENT ACCOUNTING AND FINANCIAL MODELLING (SRI LANKA)	
18	3IIS599	5	INTERACTIVE MARKETING (SRI LANKA)	
19	3ISE504	5	ALGORITHMS AND DATA STRUCTURES (SRI LANKA)	

20	3ISE510	5	NETWORK APPLICATION DEVELOPMENT (SRI LANKA)	
21	3ISE513	5	EVENT DRIVEN PROGRAMMING (SRI LANKA)	
22	3ISE514	5	OBJECT-ORIENTED DESIGN (SRI LANKA)	This module examines the techniques and methods appropriate for the design of object-oriented software. The conceptual foundations of the object-oriented approach are covered, and students will acquire practical skills in object oriented design, and also in the implementation of such designs.
23	3ISE515	5	SOFTWARE ENGINEERING GROUP PROJECT (SRI LANKA)	An extended piece of work covering a range of activities within the software engineering lifecycle will be undertaken as part of a group. This will motivate and exercise generic skills and the adoption of a professional approach to quality, management and conduct.
24	3ISE517	5	INTERNET APPLICATION PROGRAMMING (SRI LANKA)	
25	3ISE518	5	COMPUTER SYSTEMS ORGANISATION (SRI LANKA)	
26	3ISE519	5	SOFTWARE ENGINEERING (SRI LANKA)	
27	3ISE530	5	PREPARING FOR A PLACEMENT (IIT SRI LANKA)	
28	3ISE541	5	C# .Net Framework Programming (Sri Lanka)	
29	3ISE550	5	OBJECT-ORIENTED PROGRAMMING (SRI LANKA)	This module covers the use of data abstraction in software development. Students will learn how to design data abstractions, develop class hierarchies and implement them in at least one appropriate object oriented programming language.

30	3ISE630	5	PROFESSIONAL DEVELOPMENT IN COMPUTER SCIENCE (IIT SRI LANKA)	
31	3ISY509	5	PROJECT MANAGEMENT	This module aims to give students an understanding of the compleexities of project management. PM methods such as PRINCE2 & DSDM are introduced together with the main techniques for planning, monitoring and controlling the project. Concerpts such as risk assessment, quality & project roles are also introduced.
32	3ISY550	5	Computer Forensics Investigation	
33	3ISY551	5	DISTRIBUTED INFORMATION SYSTEMS	
34	3ISY552	5	REQUIREMENTS ANALYSIS	
35	3ISY553	5	DATABASE MANAGEMENT SYSTEMS	
36	3ISY555	5	RAPID APPLICATION DEVELOPMENT	
37	3ISY557	5	INTRODUCTION TO PROGRAMMING	THIS MODULE PROVIDES A COMPREHENSIVE HANDS-ON INTRODUCTION TO IMPERATIVE PROGRAMMING USING JAVA. STUDENTS WILL BE INTRODUCED TO BOTH TRADITIONAL ADN MODERN STYLE PROGRAMMING WITH A GRAPHICAL USER INTERFACE. THE MODULE IS FOCUSED ON FUNDAMENTALS OF JAVA PROGRAMMING LANGUAGE ALONG WITH HANDS-ON EXPERIENCE ON THE PROCESS OF OBJECT OIENTED ANALYSIS AND DESIGN.
38	3ISY558	5	BUSINESS ORGANISATION AND COMMUNICATION	The way business operates & the role of Information Systems within organisations. Module also reviews the approaches used for evaluating business value of Information Systems and problems involved.
39	3ISY560	5	MANAGEMENT ACCOUNTING AND FINANCIAL MODELLING	Mathematical and statistical techniques for accounting and business management. Introduction to the derivative securities market.
40	3ISY599	5	INTERACTIVE MARKETING	This module will provide the student with an introduction to marketing concepts and apply these in a practical context using web-based technologies. Students will also develop an understanding of, and apply, some of the techniques used in market research in a business

				context.
41	3MMC500	5	COMMUNICATION SIGNAL PROCESSING	To examine the characteristics of signals and processes typically found in communications systems; to introduce the theory of sinusoidal signals, modulation and filtering; to use the Fourier transform for convolution and to represent transfer functions and impulse reponses.
42	3MMC501	5	COMMUNICATIONS SYSTEMS	This module will introduce many of the fundamental ideas required for communications receiver design, digital communications, and mobile communications networks.
43	3MMC502	5	BROADCAST MEDIA SYSTEMS	Aims to show how humans perceive and physiologically react to audio and visual information; the principles of analogue and digital audio broadcasting; TV broadcasting in all formats; digital compression of signals; principles and development of video; the principles and applications of Internet broadcasting.
44	3MMC503	5	NETWORK ENGINEERING	
45	3MMC599	5	Engineering Professional Practice	To inculcate appreciation of key elements of professional engineering practice.
46	3MTS570	5	INTERACTIVE MULTIMEDIA	
47	3MTS572	5	MATHEMATICAL MODELLING OF CONTINUOUS SYSTEMS	
48	3MTS580	5	3D Computer Graphics	Introduction to communication of ideas through three dimensional computer modelling. The module covers basic modelling principles and simple animation using 3D Studio Max, an industrial standard package. Ideas are developed through a series of design methods including freehand sketching before their transfer to the computer.
49	3MTS581	5	Soft Media Processing	Introduces software techniques for capturing, processing and displaying and distributing video, audio and image data.
50	3MTS590	5	WEB DESIGN AND DEVELOPMENT	
51	3MTS592	5	HARD INTERFACES	
52	3MTS594	5	MULTIMEDIA SIMULATION	
53	3MTS595	5	INTERACTION DESIGN FOR COMMUNICATION	
54	3MTS597	5	VISUALISATION PROJECT	

55	3SFE504	5	ALGORITHMS AND DATA STRUCTURES	This module introduces the advanced data structures and algorithms that are used in various application areas of computing. Emphasis will be on the theoretical knowledge of binary search trees eg. AVL, B-trees, strings manipulation and compression, hash tables, dynamic storage allocation, garbage collection, graphs, sorting and searching. General techniques of algorithm design will be covered.
56	3SFE508	5	COMPUTER GRAPHICS	Data structures suitable for graphics, algorithms for image generation of graphic objects. The application of these techniques is described together with the nature of the devices & systems which implement computer graphics.
57	3SFE509	5	HUMAN-COMPUTER INTERFACE DESIGN	This module introduces stuents to the theoretical aspects of human-computer interaction and user interface design techniques for developing user-friendly and usable software interfaces. Tis module also investigates the use of input/output tools, visual requirements and software engineering concepts.
58	3SFE510	5	NETWORK APPLICATION DEVELOPMENT	To provide experience in the design and development of distributed systems. Two main approaches will be fully studied: the communication oriented (direct access to network interface) and problem oriented (use of remote procedure calls). Practical work will use a TCP/IP Unix based network.
59	3SFE513	5	EVENT DRIVEN PROGRAMMING	Graphical computing requires the programmer to draw output and to respond to user-generated events. The Module shows how an object-oriented approach can be used to develop such applications using provided C++ class hierarchies.
60	3SFE514	5	OBJECT-ORIENTED DESIGN	This module examines the techniques and methods appropriate for the design of object-oriented software. The conceptual foundations of the object-oriented approach are covered, and students will acquire practical skills in object oriented design, and also in the implementation of such designs.
61	3SFE515	5	SOFTWARE ENGINEERING GROUP PROJECT	An extended piece of work covering a range of activities within the software engineering lifecycle will be undertaken as part of a group. This will motivate and exercise generic skills and the adoption of a professional approach to quality, management and conduct.
62	3SFE517	5	INTERNET APPLICATION PROGRAMMING	
63	3SFE518	5	COMPUTER SYSTEMS ORGANISATION	
64	3SFE519	5	SOFTWARE ENGINEERING	
65	3SFE530	5	PREPARING FOR A PLACEMENT	THIS MODULE IS ONLY OPEN TO STUDENTS WHO ARE REGISTERED ON A SANDMICH MODE CSCS COURSE. STUDENTS MUST PASS THIS MODULE TO PROGRESS ONTO THEIR PLACEMENT. THIS MODULE PREPARES STUDENTS TO

				OBTAIN AND UNDER A WORK PLACEMENT AND CONTINUING PROFESSIONAL DEVELOPMENT.
66	3SFE540	5	Java Mobile Application Development	
67	3SFE541	5	C# .NET FRAMEWORK PROGRAMMING	This module give s the student the skills and knowledge to develop applications of the Compact Framework an open source cross-platform virtual operating system. Topics and issues covered in the module include overviews of Windows CE, .net and the Compact Framework itself. The relationship and interoperability between enterprise desktop and mobile applications is explored. Framework limitations and native coding issues in relation to telephony, networks, and messaging services, development for enterprise applications and web services will be covered. C# will be used as the main programming language for application development.
68	3SFE542	5	Mobile User Interface Development	This module introduces the practice and theory of human-computer interaction and user interface design with an emphasis on mobile devices. It covers techniques for developing user-friendly and effective graphical user interfaces within a standalone system. It also investigates the use of the task analysis methodology, visual requirements usability and evaluation techniques.
69	3SFE550	5	OBJECT-ORIENTED PROGRAMMING	This module covers the use of data abstraction in software development. Students will learn how to design data abstractions, develop class hierarchies and implement them in at least one appropriate object oriented programming language.

Appendix B2: University of Westminster SRS Year 2 data

mid	id	level	name	desc
1	3CCE632	6	REAL-TIME AND EMBEDDED SYSTEMS	The module covers the concepts of real-time and embedded systems, the techniques for their design and implementation, and hands on experience in making the software and hardware components work together in embedded system environment.
2	3CCE633	6	Embedded Processor Architectures	To give an understanding of the design, at systems-on-silicon level, of microprocessors and microcontrollerss for both general purpose and DSP applications. Particular emphasis is given to the cost and benefit of a range of approaches to performance enhancements in the light of the processors target application.
3	3ECE602	6	Advanced System Analysis and Design	To bring your understanding of the theory of linear system and signal analysis to the level that you can effectively use it for designing of control engineering an d signal processing systems in deterministic and random signals environments. To teach mathematical tools needed for understanding and practical use of this theory. To familiarise you with selected uses and adverse effects o nonlinearities in system analysis.
4	3ECE603	6	Analog Devices and Applicaitons	To enable you to analyse, design, simulate and test analogue circuits from a given specification. To give you the confidence to investigate possible solutions with reference to textbooks and your own ability to improvise and create circuit solutions.
5	3ECE607	6	School Ambassador Scheme	To provide experience in communicating technical information and skills both on an individual basis and to groups. To develop organisational interpersonal and managerial skills. To build self confidence. To provide a taste of teaching as a career. To help gain an understanding of a range of teaching methods that can be used to teach mathematical and scientific principles (as applied to engineering) to pupils at different stages of the secondary school curriculum.
6	3ECE615	6	INDUSTRIAL MANAGEMENT	To provide an understanding of production operations as a major functional area of business, in the main management decision areas of process, capacity, inventory, work force and quality. To improve the students' decision making capacity by utilising all the underlying disciplines; behavioural, quantitative, economic and systems.
7	3ECE616	6	INDIVIDUAL PROJECT	To present the challenge of: analysing, investigating feasibility, proposing a solution and solution realising under a tight specified time and cost budget for a real-world engineering problem. To develop, enhance and promote problem defintion, analysis, design, construction, measurement, evaluation, presentation and communication skills.

8	3ECE621	6	DIGITAL SIGNAL PROCESSING DESIGN	TO DEMONSTRATE DIGITAL SIGNAL PROCESSING CONCEPTS: TO RELATE CONTINOUS-TIME SIGNALS AND SYSTEMS TO DISCRETE-TIME COUNTERPARTS: TO GAIN IN-DEPTH EXPERIENCE OF DIGITAL FILTER DESIGN, TAKING ACCOUNT OF PRACTICAL IMPLEMENTATION CONSTRAINTS.
9	3ECE624	6	System Design Project	To give experience in working in a design team developing both hardware and software to produce a programmable ASIC. To enable you to implement design for test procedures and use CAE hierarchical design tools. To give you the experience in interfacing custom processors to a standard host computer and the programming of the host interface software.
10	3ECM100	6	Individual Project	To present the challenge of analysing investigating feasibility proposing a solution and solution realisation under a tightly specified time and cost budget for a real world engineering problem. To develop enhance and promote problem definition mathematical analysis, design, synthesis, construction, measurement, evaluation, presentation and communication skills. To broaden your horizons in market place product economics, competitiveness and timeliness as well as promoting industrial involvement use and development of industrially relevant tools.
11	3EDM671	6	MAJOR PROJECT	To provide an opportunity to demonstrate a student's cumulative learning through all stages of the course. To promote opportunities for subject specialisation in conjunction with other selected optional modules. To execute a major piece of self-initiated and personal work as the foundation of a professional portfolio.
12	3EDM672	6	CLIENT-BASED PROJECT	The module enables students to undertake projects defined by organisations outside the University, appropriate to their course. The student should establish the client and the project, not the University. The project should enable the student to demonstrate some or all of the course objectives and to involve approximately 120 hours of work. Each student will be allocated a project supervisor from the academic staff.
13	3EEE612	6	ANALOG MICROELECTRONICS	To enable students to analyse and design complex analog circuits where sometimes only approximate analytical methods can be applied. To qualify them to design circuits with a given frequency response. To improve their ability and skills in using design tools based on SPICE, for modelling devices and circuits. To introduce them to special design issues ie design for testability
14	3EEE613	6	ALGORITHM REALISATION	To give students experience in translating mathematical descriptions of signal processing to system realisations. More specifically to enable students to implement a transfer function described in either the Laplace domain (for continuous-time signals) or the z domain (for discrete-time variables) as an active filter, a switched capacitor filter, a hardware ASIC or a software program.
15	3IIS612	6	DISTRIBUTED BUSINESS APPLICATIONS (SRI LANKA)	This module surveys the range of end-user applications based on distributed information systems, and discusses their design & development, together with their industrial and commercial advantages (and pitfalls).

16	3IIS651	6	Designing Information Systems (Sri Lanka)	
17	3IIS652	6	Research in Information Systems (Sri Lanka)	
18	3IIS654	6	Software Quality Management (Sri Lanka)	
19	3IIS655	6	Business Systems Management Evaluation (Sri Lanka)	
20	3IIS657	6	Information Systems Development Methods (Sri Lanka)	
21	3IIS658	6	Knowledge Management (Sri Lanka)	
22	3IIS699	6	Project (Sri Lanka)	
23	3ISE601	6	OPERATING SYSTEMS DESIGN (SRI LANKA)	Provides an in-depth treatment of the design of operating systems through a substantial laboratory based study of actual operating system source code. Practical work will include modifying or replacing some modules, porting from one architecture to another and producing performance reports.
24	3ISE609	6	REAL-TIME AND EMBEDDED SYSTEMS (SRI LANKA)	Provides the student with a sound insight into the issues associated with real-time & embedded systems & the software techniques developed to address them. A thorough knowledge of the software methodologies used to address the needs of programming real-time systems will be provided.
25	3ISE611	6	Network Software Design (Sri Lanka)	(Binary/Image)
26	3ISE613	6	SECURE LANGUAGES (SRI LANKA)	The module starts from the premise that safe reliable software is more likely to be produced when written in a language with sound constructs. The module therefore explores program safety issues and the programming language features that have been introdced to address these issues. The object-orientated language EIFFEL is currently used for paractical work.
27	3ISE616	6	REQUIREMENTS ENGINEERING (SRI	

			LANKA)	
28	3ISE617	6	INTERNET APPLICATION DESIGN (SRI LANKA)	This module applies techniques and methods of object-oriented design to large-scale, robust web applications. Typical architectures for the interaction between a user and a web application are presented and implementations are constructed using current Java technologies.
29	3ISE619	6	Network Architecture (Sri Lanka)	
30	3ISE699	6	SOFTWARE ENGINEERING PROJECT (SRI LANKA)	Project undertaken only by students studying BSc(Software Engineering). The project introduces the student to detailed in-depth study of an application area and to the writing of a critical report on the work carried out. It is a significant area of the degree and exercises skills that will be important in the student's future career and which are difficult to measure in other ways.
31	3MMC600	6	DIGITAL SIGNAL PROCESSING	Aims to provide understanding of the DSP foundations of modern communications and multi-media systems; to gain basic skills in digital filter design; become experienced in the fundamentals of real-time DSP processing of audio signals.
32	3MMC601	6	COMPUTING AND NETWORKS	Aims to familiarise the student with a variety of hardware and software techniques for communication and information transfer; aims to instil an awareness of factors affecting the choice of appropriate methods.
33	3MMC606	6	INDIVIDUAL PROJECT	Aims to present the challenge of : analysising, investigating feasaibility, proposing a solution and solution realisation under a tightly specified time and cost budget for a real-world engineering problem.
34	3MTS621	6	MATHEMATICAL PROGRAMMING	Linear programming, Integer programming. Goal and multi-objective programming. Understanding advanced linear programming techniques. The analysis of of model solutions, sensitivity analysis and subsequent recommendations. Identifying and solving problems which can be modelled using discrete or multi-objective programming. The power of heuristics for solving discrete and combinational problems.
35	3MTS622	6	TIME SERIES ANALYSIS AND FORECASTING	A broad overview of the major methodologies adopted in time series analysis and forecasting, including recent developments in the field and the use of software packages : Explains the different approaches to time series analysis; the major features of time series; fitting a Box-Jenkins model to a given time series, with the aid of a suitable statistical package, by following prescribed stages, etc.
36	3MTS623	6	SEMINARS IN STATISTICS AND OPERATIONAL RESEARCH	Current topics in Operational Research (OR) and Statistics e.g., data envelopment analysis, analytical hierarchy process, neural networks, delivered in the form of seminars given at the University by lecturers, visitors and students. Occasional attendance at seminars given outside the University. Check for pre-requisitse with the Module Leader
37	3MTS626	6	GRAPHICS AND VISUALISATION	Exposition of the power of visual information, development of various modelling and rendering processes with appropriate theoretical underpinning and enablement of specific visualisation using Java3D.

38	3MTS629	6	NONLINEAR SYSTEMS	Analysis of nonlinear systems involving various techniques of approximating the normally otherwise insoluble nonlinear equations to achieve a global understanding from local behaviour : visualise non-linear behaviour and interpret such visualisation; carry out relevant computer experimentation; interpret high level mathematical argument; communicate mathematical ideas.
39	3MTS674	6	EXPERIMENTAL USABILITY	The aim of the module is to expose students to appropriate methods of applied research and approaches to practical interface design via experimentation to determine the usability of a range of information (interactive) products. The module will stress the importance of the concepts of user involvement (user centred design and participatory design), iteration and research-based design in a wide range of contemporary products. The module will consider methods for user interface evaluation, experimental design and the statistical evaluation of the experimental results and their presentation in oral and written form. The module will consider the ethical issues of experiments involving a range of subjects.
40	3MTS690	6	Mobile Gaming	Introduces mobile gaming as a mobile computing application. It covers basic game design methodology and implementation using thick client or authoring approaches. For practical work implementation may be in either depending on the student's background.
41	3MTS691	6	3D Modelling	Three dimensional computer modelling at an advanced level developing complex objects environments and animations using computer modelling. The module builds on modelling and animation skills developed using 3D studio max during the introductory module 3D computer graphics. Students will be expected to add media from other sources and models will be expected to be more detailed. 3D models and animations are exported to a wide range of other application software, such as Flash or DirectX for many different uses. The module is concerned with animated objects but is not concerned with the production of animations as a final product.
42	3MTS694	6	DESIGN FOR USER EXPERIENCE	
43	3MTS695	6	MAJOR PROJECT	
44	3MTS696	6	IPD PROJECT	Content is based on the creation of a body of project work, which represents and communicates a chosen subset of the skills methods and theoretical ideas learned throughout the course. The focus for the project should be relevant to the student's future career direction and will be based on the design and communicaiton of an information prodcut.
45	3MTS697	6	DYNAMIC SYSTEM DEVELOPMENT METHOD	

46	3MTS699	6	MATHEMATICAL SCIENCES PROJECT	Project undertaken only by students studying B. Sc. (Math. Sc.). The project introduces the student to detailed in-depth study of an application area and to the writing of a critical report on the work carried out. It is a significant part of the degree and exercises skills that will be important in the student's future career and which are difficult to measure in other ways.
47	3PMA604	6	NATURE OF MATHEMATICS	Introducing some fundamental ideas to illustrate the nature of mathematical reasoning. Rigourous approaches to formalising concepts of number, infinity, dimension and computability are investigated. Some facility with basic algebra and abstract thinking is expected.
48	3SFE601	6	OPERATING SYSTEMS DESIGN	Provides an in-depth treatment of the design of operating systems through a substantial laboratory based study of actual operating system source code. Practical work will include modifying or replacing some modules, porting from one architecture to another and producing performance reports.
49	3SFE602	6	COMPILER DESIGN TECHNIQUES	Covers theoretical and practical aspects of compiling. The theory includes the study of grammars, finite state machines and parsing algorithms. The practical side includes the analysis of a compiler for a simple block structured language. This involves using the Unix tools, lex and yacc, and the `C' programming language to amend the compiler.
50	3ISY608	6	ONLINE ANALYTICAL PROCESSING	The development of OLAP information systems using object oriented techniques. Meeting the complex processing requirements of an organisation by utilising object oriented methods. Check pre-requisites !
51	3ISY612	6	DISTRIBUTED BUSINESS APPLICATIONS	This module surveys the range of end-user applications based on distributed information systems, and discusses their design & development, together with their industrial and commercial advantages (and pitfalls).
52	3ISY651	6	DESIGNING INFORMATION SYSTEMS	The module aims to equip students to make an informed choice about the best approach to systems development for an individual system. It will enable them to evaluate the results of the requirements analysisand use them to produce a coherent systems design.
53	3ISY652	6	RESEARCH IN INFORMATION SYSTEMS	Module focuses on research and writing skills and prepares the student for carrying out work in their final year project. Topics covered are related to information systems and the associated work (assessed or unassessed) applied these skills to the topic.
54	3ISY654	6	SOFTWARE QUALITY MANAGEMENT	The module will provide students with an understanding of the importance of Software Quality Assurance (SQA) and the need to manage the software development process within the information systems project management framework. The aim is to identify and understand the procedures involved when aiming for a software quality management accreditation.
55	3ISY655	6	BUSINESS SYSTEMS MANAGEMENT & EVALUATION	This module will provide the student with an understanding of the implications of managerial decisions applied to identifying the needs for, and the development of an information system taking a strategic management perspective in relation to evaluating investment in IS/IT.

56	3ISY657	6	INFORMATION SYSTEMS DEVELOPMENT METHODOLOGIES	This module gives a historical review of methodologies for Information Systems development with the emphasis on current practices. The module also addresses the comparison frameworks needed to choose the most suitable development process across various problem and business domains.
57	3ISY658	6	KNOWLEDGE MANAGEMENT	The module aims to equip students with the knowledge and understanding to evaluate the impact of KM on organisations today. It will give them an understanding of the roots of KM and possible future developments.
58	3ISY699	6	INFORMATION SYSTEMS PROJECT	Through both semesters at Level 3. Enables students to work independently and plan their work so as to meet deadlines set. The aim is to apply what is taught and design an application.
59	3SFE605	6	CONCURRENT PROGRAMMING	Combines both theoretical and practical programming approaches to provide the skills and knowledge to be able to design and reason about concurrent programs : Understanding basic issues and benefits of concurrent programming; how to design and implement a concurrent program; in-depth look at the concurrency features of one concurrent programming language.
60	3SFE609	6	REAL-TIME AND EMBEDDED SYSTEMS	Provides the student with a sound insight into the issues associated with real-time & embedded systems & the software techniques developed to address them. A thorough knowledge of the software methodologies used to address the needs of programming real-time systems will be provided.
61	3SFE610	6	FUNCTIONAL PROGRAMMING	Introduction to the functional programming paradigm, and coverage of the use of functional techniques in both conventional & applicative languages. A modern functional programming language will be examined in detail, and students will gain extensive experience in its use.
62	3SFE611	6	NETWORK SOFTWARE DESIGN	To provide experience in the design and development of network software. The source code of an implementation of the TCP/IP suite of protocols will be available for study purposes. Students will use this to analyse network protocols, modify these and create their own.
63	3SFE613	6	SECURE LANGUAGES	The module starts from the premise that safe reliable software is more likely to be produced when written in a language with sound constructs. The module therefore explores program safety issues and the programming language features that have been introdced to address these issues. The object-orientated language EIFFEL is currently used for paractical work.
64	3SFE615	6	WEB SITE ADMINISTRATION & MAINTENANCE	This module covers the issues involved in administering an Internet Web server. Examines the strategic choices to be made in selecting a server, and gives practical experience of the mecahnisms involved in the installation, configuration, security provisions and maintenance of a web server.
65	3SFE616	6	REQUIREMENTS ENGINEERING	The lectures and tutorials will give students the opportunity to become acquainted with a range of techniques that support requirements engineering process. This will be supported by the use of industry standard tools. Students will be required to undertake additional reading and a set of topics for further study will be given out as part of the coursework.

66	3SFE617	6	INTERNET APPLICATION DESIGN	This module applies techniques and methods of object-oriented design to large-scale, robust web applications. Typical architectures for the interaction between a user and a web application are presented and implementations are constructed using current Java technologies.
67	3SFE618	6	FORMAL METHODS	Module examines the use of formal methods in system specification. A formal specification language, eg) Z, will be covered in depth with use of suitable case studies. Areas covered : design of structured specs, use of tools to support development & rigourous reasoning about specs. The strengths & weaknesses of formal methods will be critically examined.
68	3SFE619	6	NETWORK ARCHITECTURE	To learn advanced network architectures, protocols and security issues. Experience of hands-on operation of routers and other networking equipment.
69	3SFE620	6	WEB SERVICES	The module teaches web services including its architecture, features, standards such as WSDL UDDI and SOAP and implementations. It covers design, development, installation and maintenance issues of web services based applications. Students gain theoretical knowledge by learning the Service Oriented Architecture (SOA) web service models web services standards and practical skill by designing and developing web services based applications. Students learn how to integrate web services and the J2EE environment to create scalable reliable and user friendly web applications.
70	3SFE621	6	Native Application Development	This module gives the student the necessary knowledge and practical experience to develop native Symbian OS applications for a range of platforms. The module contains lectures on C++ programming and coding standards, Symbian development tools and techniques, Symbian platform components and the Symbian specific models and APIs required for multi-platform development. This module will also provide students with an introduction into native development for other popular platforms that use C++ as the main programming language.
71	3SFE622	6	Mobile and Wireless Systems Architecture	Mobile and wireless connectivity plays an increasingly dominant role in terms of providing the infrastructure for services and applications in the 21st century. This module focuses on 2nd and 3rd generation mobile cellular networks (with particular emphasis on GSM and UMTS) as well as other wireless computing technologies such as wireless LAN. The service capabilities of such networks will be discussed as well as their operation and component architecture. Technologies such as GSM, UMTS, GPRS and Wi Fi will be examined. The module discusses how appropriate choices of mobile network connectivity can be made given the constraints of budget location and device capabilities.
72	3SFE623	6	COMPUTER SECURITY AND FORENSICS	This module examines various aspects of computer security and forensics, giving a sound introduction to theoretical and practical areas. A substantial amount of work will be laboratory based involving the deployment of security tools, the hardening of operating systems and the analysis of compromised systems.
73	3SFE630	6	PROFESSIONAL DEVELOPMENT IN COMPUTER SCIENCE	THIS MODULE IS ONLY OPEN TO STUDENTS WHO ARE REGISTERED ON A SANDWICH COURSE IN CSCS. THIS MDOULE OFFERS PRACTICAL EXPERIENCE WITHIN THE COMPUTING INDUSTRY THROUGH A WORK PLACEMENT AND CAREER DEVLOPMENT THROUGH THE BCS CONTIUNING PROFESSIONAL DEVELOPMENT SCHEME.

74	3SFE699	6	SOFTWARE ENGINEERING PROJECT	Project undertaken only by students studying BSc(Software Engineering). The project introduces the student to detailed in-depth study of an application area and to the writing of a critical report on the work carried out. It is a significant area of the degree and exercises skills that will be important in the student's future career and which are difficult to measure in other ways.
75	3SMC601	6	MOBILE RADIO SYSTEMS	Aims to develop important ideas that underpin the use of mobile radio for communication such as the use of the radio spectrum, modulation techniques and spread spectrum methods.
76	3SMT602	6	SOUND AND IMAGE PROCESSING	Aims to provide the student with the fundamentals of sound and image processing, for compression, enhancement, reconstruction and synchronisation; to provide the student with the principles of error correction with regard to transmission and recording; to acquaint the student with the techniques in hardware for sound/image processing
77	3SRT601	6	REAL-TIME AND EMBEDDED SYSTEMS	The tutorials will give students the opportunity to learn how to develop software for real-time embedded computer systems. The tutorials are taking place in the electronics laboratory (Motorola lab) where the students will use special hardware (embedded microcontroller FM-400) and the support software (C compiler and MINOS operating system) which are only available in the lab. Most of the tutorials contain exercises which are part of the marked coursework As the maximum lab capacity is 20 seats, it is important that the students stick to their group allocation.
78	3SRT602	6	DSP IMPLEMENTATIONS	
79	3TSE612	6	RF AND MICROWAVE SYSTEMS	To give theoretical insight into the RF and microwave field. To enable students to use their knowledge to design an RF or microwave amplifier. To give the adequate background to enable students to analyse other subsytem circuits such as oscillators and mixers.
80	3TSE613	6	DATA COMMUNICATIONS	To provide the students with the knowledge of modern data communications systems. To study all the important sub-systems and parameters of modern terrestial digital communications networks. To introduce digital mobile cellular system architecture and operation principles.
81	3SMC602	6	CELLULAR RADIO NETWORKS	Aims to give a broad coverage of the structure, functionality and performance of networks in telecommunications based on different types of services.
82	3SMT601	6	VIDEO BROADCASTING	Aims to provide the student with a coherent foundation in understanding video broadcasting systems as they are incorporated in commercial television; to show and contrast Analog and Digital broadcast systems in terrestrial and satellite formats; to keep the student abreast with the latest broadcast standards and technology.

Appendix B3: University of Westminster SRS study automated log

27_02_09 (year 1 undergrads)
01SemaSearch:: Description: object oriented programming modules; Request string: oo programming; Time search started: Thu Feb 26 16:32:49 GMT 2009; Search term oo matches T1: object-oriented; Score: 0.22781285117006467; Search term programming matches T3: programming; Score: 0.0032670795786742006; Matched modules: ID: 29 Score: 2| ID: 69 Score: 2| ID: 62 Score: 2| ID: 24 Score: 2| ID: 22 Score: 1| ID: 60 Score: 1| ID: 15 Score: 1| ID: 67 Score: 1| ID: 59 Score: 1| ID: 37 Score: 1| ID: 28 Score: 1| ID: 21 Score: 1| ; Time results returned: Thu Feb 26 16:33:06 GMT 2009; User returned results at: Thu Feb 26 16:34:15 GMT 2009; User section: 29, 69, 62, 24, 22, 60, 15, 67, 59, 37, 28, 21,
02SemaSearch:: Description: a module teaching rules and methods of graphical user interface designand implementation; Request string: interface design; Time search started: Thu Feb 26 23:01:30 GMT 2009; Search term interface matches T3: interface; Score: 0.003962536055007237; Search term design matches T3: design; Score: 6.476616021634721E-5; Matched modules: ID: 22 Score: 3| ID: 50 Score: 3| ID: 53 Score: 3| ID: 57 Score: 2| ID: 3 Score: 2| ID: 5 Score: 2| ID: 68 Score: 1| ID: 60 Score: 1| ; Time results returned: Thu Feb 26 23:01:46 GMT 2009; User returned results at: Thu Feb 26 23:02:08 GMT 2009; User section: 53, 57, 68,
03SemaSearch:: Description: animation; Request string: computing; Time search started: Fri Feb 27 09:09:43 GMT 2009; Search term computing matches T3: interface; Score: 0.15875196561930097; Matched modules: ID: 57 Score: 1| ID: 68 Score: 1| ; Time results returned: Fri Feb 27 09:09:50 GMT 2009; User returned results at: Fri Feb 27 09:10:15 GMT 2009; User section: 57,
04SemaSearch:: Description: graphical design; Request string: graphics; Time search started: Fri Feb 27 09:11:24 GMT 2009; Search term graphics matches T3: graphics; Score: 0.0; Matched modules: ID: 48 Score: 2| ID: 56 Score: 2| ; Time results returned: Fri Feb 27 09:11:35 GMT 2009; User returned results at: Fri Feb 27 09:12:43 GMT 2009; User section: 48, 56,
05SemaSearch:: Description: html java; Request string: web design; Time search started: Fri Feb 27 09:13:49 GMT 2009; Search term web matches T2: web; Score: 4.5095134203708984E-4; Search term design matches T3: design; Score: 0.0; Matched modules: ID: 50 Score: 6| ID: 53 Score: 6| ID: 22 Score: 6| ID: 16 Score: 5| ID: 23 Score: 5| ID: 11 Score: 4| ID: 9 Score: 4| ID: 28 Score: 4| ID: 3 Score: 4| ID: 5 Score: 4| ID: 20 Score: 4| ID: 38 Score: 3| ID: 24 Score: 3| ID: 68 Score: 3| ID: 61 Score: 3| ID: 66 Score: 3| ID: 33 Score: 2| ID: 43 Score: 2| ID: 49 Score: 2| ID: 67 Score: 2| ID: 60 Score: 2| ID: 57 Score: 2| ID: 2 Score: 2| ID: 6 Score: 2| ID: 44 Score: 2| ID: 58 Score: 2| ID: 62 Score: 1| ; Time results returned: Fri Feb 27 09:14:15 GMT 2009; User returned results at: Fri Feb 27 09:16:13 GMT 2009; User section: 50, 24, 66, 62,
06SemaSearch:: Description: internet programming; Request string: java; Time search started: Fri Feb 27 10:10:37 GMT 2009; Search term java matches T3: java; Score: 0.0; Matched modules: ID: 13 Score: 5| ID: 66 Score: 3| ID: 35 Score: 3| ; Time results returned: Fri Feb 27 10:10:41 GMT 2009; User returned results at: Fri Feb 27 10:11:15 GMT 2009; User section: 66,
07SemaSearch:: Description: internet programming; Request string: internet programming; Time search started: Fri Feb 27 10:17:13 GMT 2009; Search term internet matches T3: internet; Score: 0.0; Search term programming matches T3: programming; Score: 0.0032670795786742006; Matched modules: ID: 24 Score: 5| ID: 62 Score: 3| ID: 15 Score: 1| ID: 67 Score: 1| ID: 59 Score: 1| ID: 37 Score: 1| ID: 29 Score: 1| ID: 28 Score: 1| ID: 21 Score: 1| ID: 69 Score: 1| ; Time results returned: Fri Feb 27 10:17:36 GMT 2009; User returned results at: Fri Feb 27 10:19:48 GMT 2009; User section: 24,
08SemaSearch:: Description: java programming; Request string: java; Time search started: Fri Feb 27 14:11:41 GMT 2009; Search term java matches T3: java; Score: 0.0; Matched modules: ID: 13 Score: 5| ID: 66 Score: 3| ID: 35 Score: 3| ; Time results returned: Fri Feb 27 14:11:45 GMT 2009; User returned results at:

Fri Feb 27 14:12:39 GMT 2009; User section: 66,
09SemaSearch:: Description: programming; Request string: c#; Time search started: Fri Feb 27 14:13:50 GMT 2009; Search term c# matches T0: ; Score: 0.0; Time results returned: Fri Feb 27 14:14:01 GMT 2009; User returned results at: Fri Feb 27 14:14:12 GMT 2009; User section:
10SemaSearch:: Description: programming; Request string: c# programming; Time search started: Fri Feb 27 14:14:47 GMT 2009; Search term c# matches T0: ; Score: 0.0; Search term programming matches T3: programming; Score: 0.002814578439639584; Matched modules: ID: 62 Score: 2| ID: 24 Score: 2| ID: 15 Score: 1| ID: 67 Score: 1| ID: 59 Score: 1| ID: 37 Score: 1| ID: 29 Score: 1| ID: 28 Score: 1| ID: 21 Score: 1| ID: 69 Score: 1| ; Time results returned: Fri Feb 27 14:15:12 GMT 2009; User returned results at: Fri Feb 27 14:16:47 GMT 2009; User section: 67, 28,
11SemaSearch:: Description: i want to get somw inframtion on this module; Request string: Rapid Application Dev; Time search started: Fri Feb 27 14:18:34 GMT 2009; Search term Rapid matches T3: rapid; Score: 0.0; Search term Application matches T3: application; Score: 0.005345923298426635; Search term Dev matches T2: signal; Score: 0.23201445165652046; Matched modules: ID: 14 Score: 5| ID: 28 Score: 4| ID: 36 Score: 3| ID: 67 Score: 2| ID: 24 Score: 2| ID: 62 Score: 2| ID: 41 Score: 2| ID: 3 Score: 2| ID: 20 Score: 1| ID: 58 Score: 1| ID: 66 Score: 1| ID: 6 Score: 1| ; Time results returned: Fri Feb 27 14:19:00 GMT 2009; User returned results at: Fri Feb 27 14:19:58 GMT 2009; User section: 58,
12SemaSearch:: Description: ; Request string: Database systems; Time search started: Fri Feb 27 14:21:16 GMT 2009; Search term Database matches T2: database; Score: 0.0; Search term systems matches T3: systems; Score: 2.8650471019845157E-4; Matched modules: ID: 13 Score: 10| ID: 35 Score: 6| ID: 11 Score: 4| ID: 25 Score: 4| ID: 66 Score: 3| ID: 33 Score: 2| ID: 42 Score: 2| ID: 43 Score: 2| ID: 63 Score: 2| ID: 47 Score: 1| ; Time results returned: Fri Feb 27 14:21:33 GMT 2009; User returned results at: Fri Feb 27 14:23:38 GMT 2009; User section: 35,
13SemaSearch:: Description: Interested in learning graphics ; Request string: Graphics; Time search started: Fri Feb 27 14:27:23 GMT 2009; Search term Graphics matches T3: graphics; Score: 3.3214245600697616E-4; Matched modules: ID: 48 Score: 2| ID: 56 Score: 2| ; Time results returned: Fri Feb 27 14:27:33 GMT 2009; User returned results at: Fri Feb 27 14:27:50 GMT 2009; User section: 48, 56,
14SemaSearch:: Description: php; Request string: internet; Time search started: Fri Feb 27 14:28:57 GMT 2009; Search term internet matches T3: internet; Score: 0.0; Matched modules: ID: 24 Score: 3| ID: 62 Score: 1| ; Time results returned: Fri Feb 27 14:29:06 GMT 2009; User returned results at: Fri Feb 27 14:29:49 GMT 2009; User section: 62,
15SemaSearch:: Description: javascript; Request string: internet; Time search started: Fri Feb 27 14:31:00 GMT 2009; Search term internet matches T3: internet; Score: 0.0; Matched modules: ID: 24 Score: 3| ID: 62 Score: 1| ; Time results returned: Fri Feb 27 14:31:05 GMT 2009; User returned results at: Fri Feb 27 14:31:26 GMT 2009; User section: 62,
16SemaSearch:: Description: i want to make games; Request string: game; Time search started: Fri Feb 27 14:33:40 GMT 2009; Search term game matches T3: hard; Score: 0.09102544513972874; Time results returned: Fri Feb 27 14:33:52 GMT 2009; User returned results at: Fri Feb 27 14:34:01 GMT 2009; User section:
17SemaSearch:: Description: i want to make games; Request string: 3d; Time search started: Fri Feb 27 14:34:14 GMT 2009; Search term 3d matches T3: graphics; Score: 0.12455394199262533; Matched modules: ID: 48 Score: 2| ID: 56 Score: 2| ; Time results returned: Fri Feb 27 14:34:25 GMT 2009; User returned results at: Fri Feb 27 14:35:27 GMT 2009; User section: 56,
18SemaSearch:: Description: i would like to learn silerlight .; Request string: .net ; Time search started: Fri Feb 27 14:41:51 GMT 2009; Search term .net matches T3: net; Score: 0.009879790312733273; Matched modules: ID: 28 Score: 4| ID: 67 Score: 2| ; Time results returned: Fri Feb 27 14:42:02 GMT 2009; User returned results at: Fri Feb 27 14:42:34 GMT 2009; User section: 67,

19SemaSearch:: Description: programming in C++; Request string: c++; Time search started: Fri Feb 27 14:45:08 GMT 2009; Search term c++ matches T2: web; Score: 0.07072040305201269; Matched modules: ID: 16 Score: 5| ID: 23 Score: 5| ID: 11 Score: 4| ID: 9 Score: 4| ID: 28 Score: 4| ID: 20 Score: 4| ID: 50 Score: 3| ID: 38 Score: 3| ID: 24 Score: 3| ID: 68 Score: 3| ID: 61 Score: 3| ID: 53 Score: 3| ID: 22 Score: 3| ID: 66 Score: 3| ID: 33 Score: 2| ID: 43 Score: 2| ID: 49 Score: 2| ID: 67 Score: 2| ID: 3 Score: 2| ID: 5 Score: 2| ID: 2 Score: 2| ID: 6 Score: 2| ID: 44 Score: 2| ID: 58 Score: 2| ID: 62 Score: 1| ID: 60 Score: 1| ID: 57 Score: 1| ; Time results returned: Fri Feb 27 14:45:15 GMT 2009; User returned results at: Fri Feb 27 14:47:05 GMT 2009; User section: 60,
20SemaSearch:: Description: LEARN HOW TO CREATE WEBSITES USING CODE; Request string: INTERNET PROGRAMMING; Time search started: Fri Feb 27 14:48:24 GMT 2009; Search term INTERNET matches T3: internet; Score: 0.0; Search term PROGRAMMING matches T3: programming; Score: 0.002814578439639584; Matched modules: ID: 24 Score: 5| ID: 62 Score: 3| ID: 15 Score: 1| ID: 67 Score: 1| ID: 59 Score: 1| ID: 37 Score: 1| ID: 29 Score: 1| ID: 28 Score: 1| ID: 21 Score: 1| ID: 69 Score: 1| ; Time results returned: Fri Feb 27 14:48:42 GMT 2009; User returned results at: Fri Feb 27 14:49:34 GMT 2009; User section: 62, 67,
21SemaSearch:: Description: LEARN PROGRAMMING LANGUAGES ; Request string: .NET; Time search started: Fri Feb 27 14:50:31 GMT 2009; Search term .NET matches T3: net; Score: 0.009879263396172427; Matched modules: ID: 28 Score: 4| ID: 67 Score: 2| ; Time results returned: Fri Feb 27 14:50:41 GMT 2009; User returned results at: Fri Feb 27 14:51:40 GMT 2009; User section: 67,
22SemaSearch:: Description: looking for my module (3sfe542); Request string: mobile web xml xslt; Time search started: Fri Feb 27 14:54:43 GMT 2009; Search term mobile matches T3: mobile; Score: 0.0; Search term web matches T2: web; Score: 4.533519333538698E-4; Search term xml matches T3: java; Score: 0.10061141549871838; Search term xslt matches T2: object-oriented; Score: 0.29289087751121734; Matched modules: ID: 66 Score: 9| ID: 68 Score: 6| ID: 16 Score: 5| ID: 23 Score: 5| ID: 13 Score: 5| ID: 11 Score: 4| ID: 9 Score: 4| ID: 28 Score: 4| ID: 22 Score: 4| ID: 20 Score: 4| ID: 50 Score: 3| ID: 38 Score: 3| ID: 24 Score: 3| ID: 61 Score: 3| ID: 53 Score: 3| ID: 35 Score: 3| ID: 33 Score: 2| ID: 43 Score: 2| ID: 49 Score: 2| ID: 67 Score: 2| ID: 3 Score: 2| ID: 60 Score: 2| ID: 5 Score: 2| ID: 2 Score: 2| ID: 6 Score: 2| ID: 44 Score: 2| ID: 58 Score: 2| ID: 62 Score: 1| ID: 57 Score: 1| ID: 29 Score: 1| ID: 69 Score: 1| ; Time results returned: Fri Feb 27 14:55:19 GMT 2009; User returned results at: Fri Feb 27 14:56:23 GMT 2009; User section: 68,
11_03_09 (year 2 undergrads)
23SemaSearch:: Description: applications; Request string: internet; Time search started: Wed Mar 11 15:17:12 GMT 2009; Search term internet matches T3: internet; Score: 1.8921454814797552E-4; Matched modules: ID: 28 Score: 3| ID: 66 Score: 1| ; Time results returned: Wed Mar 11 15:17:28 GMT 2009; User returned results at: Wed Mar 11 15:18:07 GMT 2009; User section: 28, 66,
24SemaSearch:: Description: programming language; Request string: ada; Time search started: Wed Mar 11 15:18:47 GMT 2009; Search term ada matches T3: formal; Score: 0.29280961341137735; Matched modules: ID: 67 Score: 2| ; Time results returned: Wed Mar 11 15:19:02 GMT 2009; User returned results at: Wed Mar 11 15:19:23 GMT 2009; User section:
25SemaSearch:: Description: programming language; Request string: occam; Time search started: Wed Mar 11 15:19:53 GMT 2009; Search term occam matches T2: mathematical; Score: 0.5844744313204395; Matched modules: ID: 34 Score: 2| ID: 46 Score: 1| ID: 14 Score: 1| ; Time results returned: Wed Mar 11 15:20:06 GMT 2009; User returned results at: Wed Mar 11 15:20:35 GMT 2009; User section:
26SemaSearch:: Description: ; Request string: business; Time search started: Wed Mar 11 15:20:09 GMT 2009; Search term business matches T3: business; Score: 0.0; Matched modules: ID: 19 Score: 4| ID: 15 Score: 3| ID: 55 Score: 2| ID: 51 Score: 1| ; Time results returned: Wed Mar 11 15:20:18 GMT 2009; User returned results at: Wed Mar 11 15:21:31 GMT 2009; User section: 15, 55, 51,
27SemaSearch:: Description: I would like to know more about web design; Request string: Web Design; Time search started: Wed Mar 11 15:24:21 GMT 2009;

Search term Web matches T2: web; Score: 2.2600871600813292E-4; Search term Design matches T3: design; Score: 5.839808329168686E-4; Matched modules: ID: 42 Score: 6| ID: 25 Score: 6| ID: 23 Score: 6| ID: 62 Score: 6| ID: 49 Score: 6| ID: 9 Score: 6| ID: 48 Score: 6| ID: 8 Score: 6| ID: 20 Score: 4| ID: 56 Score: 4| ID: 28 Score: 4| ID: 66 Score: 4| ID: 3 Score: 4| ID: 45 Score: 4| ID: 30 Score: 3| ID: 74 Score: 3| ID: 76 Score: 3| ID: 70 Score: 3| ID: 64 Score: 2| ID: 50 Score: 2| ID: 52 Score: 2| ID: 17 Score: 2| ID: 16 Score: 2| ID: 53 Score: 2| ID: 58 Score: 2| ID: 35 Score: 2| ID: 68 Score: 2| ID: 29 Score: 2| ID: 7 Score: 2| ID: 43 Score: 2| ID: 11 Score: 2| ID: 33 Score: 2| ID: 10 Score: 2| ID: 69 Score: 1| ID: 18 Score: 1| ID: 54 Score: 1| ID: 80 Score: 1| ID: 73 Score: 1| ID: 44 Score: 1| ID: 22 Score: 1| ID: 12 Score: 1| ; Time results returned: Wed Mar 11 15:24:53 GMT 2009; User returned results at: Wed Mar 11 15:27:21 GMT 2009; User section: 42, 28, 66, 64, 69,
28SemaSearch:: Description: i would like to know more about php and perl; Request string: internet programming; Time search started: Wed Mar 11 15:26:20 GMT 2009; Search term internet matches T3: internet; Score: 1.8921454814797552E-4; Search term programming matches T3: programming; Score: 0.0027605925307732427; Matched modules: ID: 28 Score: 3| ID: 34 Score: 2| ID: 59 Score: 2| ID: 61 Score: 2| ID: 66 Score: 1| ID: 40 Score: 1| ; Time results returned: Wed Mar 11 15:26:38 GMT 2009; User returned results at: Wed Mar 11 15:27:31 GMT 2009; User section: 28,
29SemaSearch:: Description: ; Request string: graphics; Time search started: Wed Mar 11 15:27:42 GMT 2009; Search term graphics matches T3: graphics; Score: 0.0; Matched modules: ID: 37 Score: 1| ; Time results returned: Wed Mar 11 15:27:53 GMT 2009; User returned results at: Wed Mar 11 15:28:11 GMT 2009; User section: 37,
30SemaSearch:: Description: hello...; Request string: hello; Time search started: Wed Mar 11 15:30:24 GMT 2009; Search term hello matches T3: sound; Score: 0.14409951350955083; Matched modules: ID: 76 Score: 3| ; Time results returned: Wed Mar 11 15:30:35 GMT 2009; User returned results at: Wed Mar 11 15:30:49 GMT 2009; User section: 76,
31SemaSearch:: Description: creating web pages; Request string: web design; Time search started: Wed Mar 11 15:29:35 GMT 2009; Search term web matches T2: web; Score: 2.2600871600813292E-4; Search term design matches T3: design; Score: 5.839808329168686E-4; Matched modules: ID: 42 Score: 6| ID: 25 Score: 6| ID: 23 Score: 6| ID: 62 Score: 6| ID: 49 Score: 6| ID: 9 Score: 6| ID: 48 Score: 6| ID: 8 Score: 6| ID: 20 Score: 4| ID: 56 Score: 4| ID: 28 Score: 4| ID: 66 Score: 4| ID: 3 Score: 4| ID: 45 Score: 4| ID: 30 Score: 3| ID: 74 Score: 3| ID: 76 Score: 3| ID: 70 Score: 3| ID: 64 Score: 2| ID: 50 Score: 2| ID: 52 Score: 2| ID: 17 Score: 2| ID: 16 Score: 2| ID: 53 Score: 2| ID: 58 Score: 2| ID: 35 Score: 2| ID: 68 Score: 2| ID: 29 Score: 2| ID: 7 Score: 2| ID: 43 Score: 2| ID: 11 Score: 2| ID: 33 Score: 2| ID: 10 Score: 2| ID: 69 Score: 1| ID: 18 Score: 1| ID: 54 Score: 1| ID: 80 Score: 1| ID: 73 Score: 1| ID: 44 Score: 1| ID: 22 Score: 1| ID: 12 Score: 1| ; Time results returned: Wed Mar 11 15:29:57 GMT 2009; User returned results at: Wed Mar 11 15:31:29 GMT 2009; User section: 42, 62, 9, 56, 66, 64,
32SemaSearch:: Description: web designing; Request string: web designing; Time search started: Wed Mar 11 15:37:56 GMT 2009; Search term web matches T2: web; Score: 2.2600871600813292E-4; Search term designing matches T2: usability; Score: 0.1890507486866861; Matched modules: ID: 20 Score: 4| ID: 56 Score: 4| ID: 45 Score: 4| ID: 42 Score: 3| ID: 25 Score: 3| ID: 23 Score: 3| ID: 62 Score: 3| ID: 49 Score: 3| ID: 9 Score: 3| ID: 48 Score: 3| ID: 8 Score: 3| ID: 30 Score: 3| ID: 74 Score: 3| ID: 76 Score: 3| ID: 70 Score: 3| ID: 64 Score: 2| ID: 50 Score: 2| ID: 52 Score: 2| ID: 17 Score: 2| ID: 16 Score: 2| ID: 53 Score: 2| ID: 58 Score: 2| ID: 35 Score: 2| ID: 28 Score: 2| ID: 66 Score: 2| ID: 3 Score: 2| ID: 68 Score: 2| ID: 29 Score: 2| ID: 7 Score: 2| ID: 43 Score: 2| ID: 11 Score: 2| ID: 33 Score: 2| ID: 10 Score: 2| ID: 39 Score: 2| ID: 69 Score: 1| ID: 18 Score: 1| ID: 54 Score: 1| ID: 80 Score: 1| ID: 73 Score: 1| ID: 44 Score: 1| ID: 22 Score: 1| ID: 12 Score: 1| ; Time results returned: Wed Mar 11 15:38:11 GMT 2009; User returned results at: Wed Mar 11 15:39:06 GMT 2009; User section: 9,

33SemaSearch:: Description: how to make web pages; Request string: web design; Time search started: Wed Mar 11 16:11:41 GMT 2009; Search term web matches T2: web; Score: 2.2600871600813292E-4; Search term design matches T3: design; Score: 5.839808329168686E-4; Matched modules: ID: 42 Score: 6| ID: 25 Score: 6| ID: 23 Score: 6| ID: 62 Score: 6| ID: 49 Score: 6| ID: 9 Score: 6| ID: 48 Score: 6| ID: 8 Score: 6| ID: 20 Score: 4| ID: 56 Score: 4| ID: 28 Score: 4| ID: 66 Score: 4| ID: 3 Score: 4| ID: 45 Score: 4| ID: 30 Score: 3| ID: 74 Score: 3| ID: 76 Score: 3| ID: 70 Score: 3| ID: 64 Score: 2| ID: 50 Score: 2| ID: 52 Score: 2| ID: 17 Score: 2| ID: 16 Score: 2| ID: 53 Score: 2| ID: 58 Score: 2| ID: 35 Score: 2| ID: 68 Score: 2| ID: 29 Score: 2| ID: 7 Score: 2| ID: 43 Score: 2| ID: 11 Score: 2| ID: 33 Score: 2| ID: 10 Score: 2| ID: 69 Score: 1| ID: 18 Score: 1| ID: 54 Score: 1| ID: 80 Score: 1| ID: 73 Score: 1| ID: 44 Score: 1| ID: 22 Score: 1| ID: 12 Score: 1| ; Time results returned: Wed Mar 11 16:11:56 GMT 2009; User returned results at: Wed Mar 11 16:14:34 GMT 2009; User section: 62, 66, 69,
12_03_09 (year 2 undergrads)
34SemaSearch:: Description: ; Request string: website administration; Time search started: Wed Mar 04 13:59:16 GMT 2009; Search term website matches T3: time; Score: 0.07831703989896671; Search term administration matches T3: administration; Score: 0.005978748297005244; Matched modules: ID: 35 Score: 2| ID: 64 Score: 2| ; Time results returned: Wed Mar 04 13:59:40 GMT 2009; User returned results at: Wed Mar 04 13:59:48 GMT 2009; User section: 64,
35SemaSearch:: Description: i would like to learn more about flash; Request string: learn flash; Time search started: Thu Mar 05 09:10:03 GMT 2009; Search term learn matches T3: user; Score: 0.07857051263048655; Search term flash matches T3: data; Score: 0.10792302281220566; Matched modules: ID: 42 Score: 3| ID: 80 Score: 1| ; Time results returned: Thu Mar 05 09:10:32 GMT 2009; User returned results at: Thu Mar 05 09:11:05 GMT 2009; User section: 42,
36SemaSearch:: Description: ; Request string: programming; Time search started: Thu Mar 05 14:04:17 GMT 2009; Search term programming matches T3: programming; Score: 0.0032352117595418135; Matched modules: ID: 34 Score: 2| ID: 59 Score: 2| ID: 61 Score: 2| ID: 40 Score: 1| ; Time results returned: Thu Mar 05 14:04:28 GMT 2009; User returned results at: Thu Mar 05 14:05:40 GMT 2009; User section: 40,
37SemaSearch:: Description: even though it is boring!; Request string: networks; Time search started: Thu Mar 05 14:06:40 GMT 2009; Search term networks matches T3: networks; Score: 0.0; Matched modules: ID: 32 Score: 2| ID: 81 Score: 1| ; Time results returned: Thu Mar 05 14:06:48 GMT 2009; User returned results at: Thu Mar 05 14:07:00 GMT 2009; User section: 81,
38SemaSearch:: Description: a platform games; Request string: java games; Time search started: Thu Mar 05 14:16:45 GMT 2009; Search term java matches T3: applications; Score: 0.13482657091428008; Search term games matches T3: software; Score: 0.06411618867901744; Matched modules: ID: 25 Score: 3| ID: 30 Score: 3| ID: 62 Score: 3| ID: 74 Score: 3| ID: 27 Score: 2| ID: 65 Score: 2| ID: 15 Score: 1| ID: 51 Score: 1| ID: 18 Score: 1| ID: 54 Score: 1| ; Time results returned: Thu Mar 05 14:17:13 GMT 2009; User returned results at: Thu Mar 05 14:19:39 GMT 2009; User section:
49SemaSearch:: Description: i would like to learn how to create a professional looking web site; Request string: html; Time search started: Thu Mar 05 14:24:24 GMT 2009; Search term html matches T3: user; Score: 0.08977680623505711; Matched modules: ID: 42 Score: 3| ; Time results returned: Thu Mar 05 14:24:37 GMT 2009; User returned results at: Thu Mar 05 14:25:12 GMT 2009; User section: 42,
40SemaSearch:: Description: javascript programme; Request string: java; Time search started: Thu Mar 05 14:26:28 GMT 2009; Search term java matches T3: applications; Score: 0.13482657091428008; Matched modules: ID: 27 Score: 2| ID: 65 Score: 2| ID: 15 Score: 1| ID: 51 Score: 1| ; Time results returned: Thu Mar 05 14:26:35 GMT 2009; User returned results at: Thu Mar 05 14:27:25 GMT 2009; User section:
41SemaSearch:: Description: help on actionscript; Request string: flash design; Time search started: Thu Mar 05 14:30:44 GMT 2009; Search term flash matches T3: data; Score: 0.10792302281220566; Search term design matches T3: design; Score: 0.0; Matched modules: ID: 62 Score: 3| ID: 49 Score: 3| ID: 48 Score: 3| ID: 42 Score: 3| ID: 25 Score: 3| ID: 23 Score: 3| ID: 9 Score: 3| ID: 8 Score: 3| ID: 3 Score: 2| ID: 28 Score: 2| ID: 66 Score: 2| ID: 80 Score: 1| ; Time results returned: Thu Mar 05 14:30:59 GMT 2009; User returned results at: Thu Mar 05 14:32:35 GMT 2009; User section: 42, 9, 28,

42SemaSearch:: Description: i like to learn more about netwoks; Request string: network; Time search started: Thu Mar 05 14:37:37 GMT 2009; Search term network matches T3: network; Score: 8.217442848901904E-5; Matched modules: ID: 25 Score: 3| ID: 62 Score: 3| ID: 29 Score: 2| ID: 68 Score: 2| ; Time results returned: Thu Mar 05 14:37:50 GMT 2009; User returned results at: Thu Mar 05 14:39:09 GMT 2009; User section: 25,
43SemaSearch:: Description: graphic design; Request string: graphic design; Time search started: Thu Mar 12 11:39:22 GMT 2009; Search term graphic matches T3: programming; Score: 0.13729591111704326; Search term design matches T3: design; Score: 6.439114970235934E-5; Matched modules: ID: 62 Score: 3| ID: 49 Score: 3| ID: 48 Score: 3| ID: 42 Score: 3| ID: 25 Score: 3| ID: 23 Score: 3| ID: 9 Score: 3| ID: 8 Score: 3| ID: 34 Score: 2| ID: 59 Score: 2| ID: 61 Score: 2| ID: 3 Score: 2| ID: 28 Score: 2| ID: 66 Score: 2| ID: 40 Score: 1| ; Time results returned: Thu Mar 12 11:39:39 GMT 2009; User returned results at: Thu Mar 12 11:40:04 GMT 2009; User section: 40,
44SemaSearch:: Description: I would like to learn more about database and its design; Request string: database design; Time search started: Thu Mar 12 11:41:27 GMT 2009; Search term database matches T3: applications; Score: 0.10976575838670757; Search term design matches T3: design; Score: 6.439114970235934E-5; Matched modules: ID: 62 Score: 3| ID: 49 Score: 3| ID: 48 Score: 3| ID: 42 Score: 3| ID: 25 Score: 3| ID: 23 Score: 3| ID: 9 Score: 3| ID: 8 Score: 3| ID: 27 Score: 2| ID: 65 Score: 2| ID: 3 Score: 2| ID: 28 Score: 2| ID: 66 Score: 2| ID: 15 Score: 1| ID: 51 Score: 1| ; Time results returned: Thu Mar 12 11:41:47 GMT 2009; User returned results at: Thu Mar 12 11:43:23 GMT 2009; User section: 62, 48, 66,
45SemaSearch:: Description: learn more about the java; Request string: java; Time search started: Thu Mar 12 11:44:19 GMT 2009; Search term java matches T3: applications; Score: 0.13447934731642608; Matched modules: ID: 27 Score: 2| ID: 65 Score: 2| ID: 15 Score: 1| ID: 51 Score: 1| ; Time results returned: Thu Mar 12 11:44:32 GMT 2009; User returned results at: Thu Mar 12 11:45:03 GMT 2009; User section:
46SemaSearch:: Description: techniques on creating web pages; Request string: web design; Time search started: Thu Mar 12 11:48:28 GMT 2009; Search term web matches T2: web; Score: 4.533519333538698E-4; Search term design matches T3: design; Score: 6.439114970235934E-5; Matched modules: ID: 42 Score: 6| ID: 25 Score: 6| ID: 23 Score: 6| ID: 62 Score: 6| ID: 49 Score: 6| ID: 9 Score: 6| ID: 48 Score: 6| ID: 8 Score: 6| ID: 20 Score: 4| ID: 56 Score: 4| ID: 28 Score: 4| ID: 66 Score: 4| ID: 3 Score: 4| ID: 45 Score: 4| ID: 30 Score: 3| ID: 74 Score: 3| ID: 76 Score: 3| ID: 70 Score: 3| ID: 64 Score: 2| ID: 50 Score: 2| ID: 52 Score: 2| ID: 17 Score: 2| ID: 16 Score: 2| ID: 53 Score: 2| ID: 58 Score: 2| ID: 35 Score: 2| ID: 68 Score: 2| ID: 29 Score: 2| ID: 7 Score: 2| ID: 43 Score: 2| ID: 11 Score: 2| ID: 33 Score: 2| ID: 10 Score: 2| ID: 69 Score: 1| ID: 18 Score: 1| ID: 54 Score: 1| ID: 80 Score: 1| ID: 73 Score: 1| ID: 44 Score: 1| ID: 22 Score: 1| ID: 12 Score: 1| ; Time results returned: Thu Mar 12 11:48:43 GMT 2009; User returned results at: Thu Mar 12 11:50:05 GMT 2009; User section: 42, 28,
16/17_03_09 (year 2 undergrads)
47SemaSearch:: Description: i'd like to learn more about op sys design; Request string: java; Time search started: Mon Mar 16 14:50:49 GMT 2009; Search term java matches T3: applications; Score: 0.13413423699627972; Matched modules: ID: 27 Score: 2| ID: 65 Score: 2| ID: 15 Score: 1| ID: 51 Score: 1| ; Time results returned: Mon Mar 16 14:50:56 GMT 2009; User returned results at: Mon Mar 16 14:51:02 GMT 2009; User section: 27,
48SemaSearch:: Description: learn more about 3d design; Request string: 3d design; Time search started: Mon Mar 16 14:52:34 GMT 2009; Search term 3d matches T2: gaming; Score: 0.15507146360118745; Search term design matches T3: design; Score: 6.429807100195666E-5; Matched modules: ID: 76 Score: 3| ID: 62 Score: 3| ID: 49 Score: 3| ID: 48 Score: 3| ID: 42 Score: 3| ID: 25 Score: 3| ID: 23 Score: 3| ID: 9 Score: 3| ID: 8 Score: 3| ID: 61 Score: 2| ID: 59 Score: 2| ID: 34 Score: 2| ID: 4 Score: 2| ID: 32 Score: 2| ID: 60 Score: 2| ID: 24 Score: 2| ID: 77 Score: 2| ID: 1 Score: 2| ID: 3 Score: 2| ID: 28 Score: 2| ID: 66 Score: 2| ID: 40 Score: 1| ID: 63 Score: 1| ID: 26 Score: 1| ID: 81 Score: 1| ; Time results returned: Mon Mar 16 14:52:58 GMT 2009; User returned results at: Mon Mar 16 14:54:38 GMT 2009; User section:

49SemaSearch:: Description: I'm interested in databases; Request string: database systems; Time search started: Tue Mar 17 11:52:40 GMT 2009; Search term database matches T3: applications; Score: 0.10954397668126656; Search term systems matches T3: systems; Score: 1.4038671244824284E-4; Matched modules: ID: 56 Score: 4| ID: 20 Score: 4| ID: 48 Score: 3| ID: 23 Score: 3| ID: 27 Score: 2| ID: 65 Score: 2| ID: 1 Score: 2| ID: 77 Score: 2| ID: 71 Score: 2| ID: 60 Score: 2| ID: 58 Score: 2| ID: 55 Score: 2| ID: 53 Score: 2| ID: 52 Score: 2| ID: 38 Score: 2| ID: 24 Score: 2| ID: 19 Score: 2| ID: 17 Score: 2| ID: 16 Score: 2| ID: 15 Score: 1| ID: 51 Score: 1| ID: 75 Score: 1| ID: 79 Score: 1| ; Time results returned: Tue Mar 17 11:53:09 GMT 2009; User returned results at: Tue Mar 17 11:55:28 GMT 2009; User section: 56, 20, 58, 52,
50SemaSearch:: Description: CCNA; Request string: networks; Time search started: Tue Mar 17 11:58:08 GMT 2009; Search term networks matches T3: networks; Score: 0.0; Matched modules: ID: 32 Score: 2| ID: 81 Score: 1| ; Time results returned: Tue Mar 17 11:58:17 GMT 2009; User returned results at: Tue Mar 17 11:58:38 GMT 2009; User section: 32,
51SemaSearch:: Description: 3d, animation; Request string: computer graphics; Time search started: Tue Mar 17 12:01:24 GMT 2009; Search term computer matches T3: computer; Score: 0.0; Search term graphics matches T3: graphics; Score: 0.0; Matched modules: ID: 72 Score: 3| ID: 73 Score: 3| ID: 37 Score: 1| ; Time results returned: Tue Mar 17 12:01:50 GMT 2009; User returned results at: Tue Mar 17 12:04:04 GMT 2009; User section: 37,
52SemaSearch:: Description: designing games with multiplayer feature; Request string: graphics 3d multi; Time search started: Tue Mar 17 12:05:12 GMT 2009; Search term graphics matches T3: graphics; Score: 0.0; Search term 3d matches T2: gaming; Score: 0.15507146360118745; Search term multi matches T3: applications; Score: 0.11324951171979467; Matched modules: ID: 76 Score: 3| ID: 61 Score: 2| ID: 59 Score: 2| ID: 34 Score: 2| ID: 4 Score: 2| ID: 32 Score: 2| ID: 60 Score: 2| ID: 24 Score: 2| ID: 77 Score: 2| ID: 1 Score: 2| ID: 27 Score: 2| ID: 65 Score: 2| ID: 37 Score: 1| ID: 40 Score: 1| ID: 63 Score: 1| ID: 26 Score: 1| ID: 81 Score: 1| ID: 15 Score: 1| ID: 51 Score: 1| ; Time results returned: Tue Mar 17 12:05:40 GMT 2009; User returned results at: Tue Mar 17 12:07:34 GMT 2009; User section: 37,

Appendix B4: University of Westminster SRS study Year 1 log analyses

University of Westminster SRS study Year 1 interpolated average Precision Recall											
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
SemaCS no personalisation	0.6684	0.6684	0.6684	0.6207	0.5686	0.5686	0.4370	0.4333	0.3815	0.3815	0.3815
SemaCS with personalisation	0.6538	0.6538	0.6538	0.6380	0.6348	0.5418	0.3404	0.3420	0.2928	0.2655	0.2665
UoW SRS search	0.6667	0.6667	0.6667	0.5714	0.4762	0.4762	0.3333	0.3333	0.2857	0.2857	0.2857

University of Westminster SRS Year 1 average				
	Recall	**Precision**	**MRR**	**F-score**
SemaCS no personalisation	0.646032	0.424979	0.666667	0.512693
SemaCS with personalisation	0.669048	0.472677	0.619048	0.553975
UoW SRS search	0.435714	0.666667	0.666667	0.526998

Appendix B4 A: SemaCS (no personalisation) Year 1 P/R and F-score

Q-ID	Student Query description	Student Query	Expert generated query matches (ID)	SemaCS generated (no personalisation)	matched	mcount	count	recall	precision	f-score
1	object oriented programming modules	oo programming	3ISY557, 3SFE513, 3SFE517, 3SFE541, 3SFE550, [3ISE550, 3ISE517, 3ISE541, 3IIS557, 3ISE513]	2:[3ISE550, 3SFE550, 3SFE517, 3ISE517] 1:[3ISE514, 3SFE514, 3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE541, 3ISE513]	10	10	12	1	0.833333333	0.909090909
2	a module teaching rules and methods of graphical user interface designand implementation	interface design	3SFE509, [3SFE542, 3ISY557, 3IIS557]	3:[3ISE514, 3MTS590, 3MTS595] 2:[3SFE509, 3ECE515, 3ECE517] 1:[3SFE542, 3SFE514]	2	4	8	0.5	0.25	0.333333333
3	animation	computing	none	1:[3SFE509, 3SFE542]	0	0	2	0	0	0
4	graphical design	graphics	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2:[3MTS580, 3SFE508]	2	4	2	0.5	1	0.666666667
5	html java	web design	3SFE517, [3ISE517, 3MTS590]	6:[3MTS590, 3MTS595, 3ISE514] 5:[3IIS558, 3ISE515] 4:[3IIS551, 3IIS503, 3ISE541, 3ECE515, 3ECE517, 3ISE510] 3:[3ISY558, 3ISE517, 3SFE542, 3SFE515,3SFE540] 2:[3ISY551, 3MMC502, 3MTS581, 3SFE541, 3SFE514, 3SFE509, 3ECE514, 3ECE521, 3MMC503, 3SFE510] 1:[3SFE517]	3	3	27	1	0.111111111	0.2
6	internet programming	java	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	5:[3IIS553] 3:[3SFE540, 3ISY553]	1	5	3	0.2	0.333333333	0.25
7	internet programming	internet programming	3SFE517, [3ISE517]	5:[3ISE517] 3:[3SFE517] 1:[3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE550, 3ISE541, 3ISE513, 3SFE550]	2	2	10	1	0.2	0.333333333
8	java programming	java	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	5:[3IIS553] 3:[3SFE540, 3ISY553]	1	5	3	0.2	0.333333333	0.25
9	programming	c#	3SFE541, [3ISE541]	none	0	2	0	0	0	0
10	programming	c# programming	3SFE541, [3ISE541]	2:[3SFE517, 3ISE517] 1:[3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE550, 3ISE541, 3ISE513, 3SFE550]	2	2	10	1	0.2	0.333333333

11	i want to get somw inframtion on this module	Rapid Applicati on Dev	3ISY555, [3IIS555]	5:[3IIS555] 4:[3ISE541] 3:[3ISY555] 2:[3SFE541, 3ISE517, 3SFE517, 3MMC500, 3ECE515] 1:[3ISE510, 3SFE510, 3SFE540, 3ECE521]	2	2	12	1	0.166666667	0.285714286
12		Databas e systems	3ISY553, [3IIS553]	10:[3IIS553] 6:[3ISY553] 4:[3IIS551, 3ISE518] 3:[3SFE540] 2:[3ISY551, 3MMC501, 3MMC502, 3SFE518] 1:[3MTS572]	2	2	10	1	0.2	0.333333333
13	Interested in learning graphics	Graphics	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2:[3MTS580, 3SFE508]	2	4	2	0.5	1	0.666666667
14	php	internet	3SFE517, [3ISE517, 3MMC502]	3:[3ISE517] 1:[3SFE517]	2	3	2	0.666666667	1	0.8
15	i want to make games	game	none	none	0	0	0	0	0	0
16	i want to make games	3d	3SFE508, 3MTS580	2:[3MTS580, 3SFE508]	2	2	2	1	1	1
17	i would like to learn silerlight .	.net	3SFE541, [3ISE541]	4:[3ISE541] 2:[3SFE541]	2	2	2	1	1	1
18	programming in C++	c++	3SFE504, [3ISE504]	5:[3IIS558, 3ISE515] 4:[3IIS551, 3IIS503, 3ISE541, 3ISE510] 3:[3MTS590, 3ISY558, 3ISE517, 3SFE542, 3SFE515, 3MTS595, 3ISE514, 3SFE540] 2:[3ISY551, 3MMC502, 3MTS581, 3SFE541, 3ECE515,3ECE517, 3ECE514, 3ECE521, 3MMC503, 3SFE510] 1:[3SFE517, 3SFE514, 3SFE509]	0	2	27	0	0	0
19	LEARN HOW TO CREATE WEBSITES USING CODE	INTERN ET PROGR AMMIN G	3SFE517, [3ISE517]	5:[3ISE517] 3:[3SFE517] 1:[3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE550, 3ISE541, 3ISE513, 3SFE550]	2	2	10	1	0.2	0.333333333
20	LEARN PROGRAMMING LANGUAGES	.NET	3SFE541, [3ISE541]	4:[3ISE541] 2:[3SFE541]	2	2	2	1	1	1
21	looking for my module (3sfe542)	mobile web xml xslt	[3SFE541, 3ISE541, 3SFE542]	9:[3SFE540] 6:[3SFE542] 5:[3IIS558, 3ISE515, 3IIS553] 4:[3IIS551, 3IIS503, 3ISE541, 3ISE514, 3ISE510] 3:[3MTS590, 3ISY558, 3ISE517, 3SFE515, 3MTS595, 3ISY553] 2:[3ISY551, 3MMC502, 3MTS581, 3SFE541, 3ECE515, 3SFE514,3ECE517, 3ECE514, 3ECE521, 3MMC503, 3SFE510] 1:[3SFE517, 3SFE509, 3ISE550, 3SFE550]	3	3	31	1	0.096774194	0.176470588

Appendix B4 B: SemaCS (no personalisation) Year 1 interpolated average P/R and MRR

Study 2 SemaCS search (no personalisation) Recall Precision and MRR									
Query №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/ MRR
1	3ISE550, 3SFE550, 3SFE517, 3ISE517, 3ISE514, 3SFE514, 3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE541, 3ISE513	12	10	3ISY557, 3SFE513, 3SFE517, 3SFE541, 3SFE550, [3ISE550, 3ISE517, 3ISE541, 3IIS557, 3ISE513]	10	1,2,3,4,5,7,9,10,11,12	1.000000	0.833333	1.000000
2	3ISE514, 3MTS590, 3MTS595, 3SFE509, 3ECE515, 3ECE517, 3SFE542, 3SFE514	8	4	3SFE509, [3SFE542, 3ISY557, 3IIS557]	2	4,7	0.500000	0.250000	0.250000
3	3SFE509, 3SFE542	2	0	none	0	0	0.000000	0.000000	0.000000
4	3MTS580, 3SFE508	2	4	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2	1,2	0.500000	1.000000	1.000000
5	3MTS590, 3MTS595, 3ISE514, 3IIS558, 3ISE515, 3IIS551, 3IIS503, 3ISE541, 3ECE515, 3ECE517, 3ISE510, 3ISY558, 3ISE517, 3SFE542, 3SFE515,3SFE540, 3ISY551, 3MMC502, 3MTS581, 3SFE541, 3SFE514, 3SFE509, 3ECE514, 3ECE521, 3MMC503, 3SFE510, 3SFE517	27	3	3SFE517, [3ISE517, 3MTS590]	3	1,13,27	1.000000	0.111111	1.000000
6	3IIS553, 3SFE540, 3ISY553	3	5	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	1	2	0.200000	0.333333	0.500000
7	3ISE517, 3SFE517, 3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE550, 3ISE541, 3ISE513, 3SFE550	10	2	3SFE517, [3ISE517]	2	1,2	1.000000	0.200000	1.000000
8	5:[3IIS553] 3:[3SFE540, 3ISY553]	3	5	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	1	2	0.200000	0.333333	0.500000
9	none	0	2	3SFE541, [3ISE541]	0	0	0.000000	0.000000	0.000000
10	3SFE517, 3ISE517, 3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE550, 3ISE541, 3ISE513, 3SFE550	10	2	3SFE541, [3ISE541]	2	4,8	1.000000	0.200000	0.250000
11	3IIS555, 3ISE541, 3ISY555, 3SFE541, 3ISE517, 3SFE517, 3MMC500, 3ECE515, 3ISE510, 3SFE510, 3SFE540, 3ECE521	12	2	3ISY555, [3IIS555]	2	1,3	1.000000	0.166667	1.000000
12	3IIS553, 3ISY553, 3IIS551, 3ISE518, 3SFE540, 3ISY551, 3MMC501, 3MMC502, 3SFE518, 3MTS572	10	2	3ISY553, [3IIS553]	2	1,2	1.000000	0.200000	1.000000
13	3MTS580, 3SFE508	2	4	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2	1,2	0.500000	1.000000	1.000000
14	3ISE517, 3SFE517	2	3	3SFE517, [3ISE517, 3MMC502]	2	1,2	0.666667	1.000000	1.000000

15	none	0	0	none	0	0	0.000000	0.000000	0.000000
16	3MTS580, 3SFE508	2	2	3SFE508, 3MTS580	2	1,2	1.000000	1.000000	1.000000
17	3ISE541, 3SFE541	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
18	3IIS558, 3ISE515, 3IIS551, 3IIS503, 3ISE541, 3ISE510, 3MTS590, 3ISY558, 3ISE517, 3SFE542, 3SFE515, 3MTS595, 3ISE514, 3SFE540, 3ISY551, 3MMC502, 3MTS581, 3SFE541, 3ECE515,3ECE517, 3ECE514, 3ECE521, 3MMC503, 3SFE510, 3SFE517, 3SFE514, 3SFE509	27	2	3SFE504, [3ISE504]	0	0	0.000000	0.000000	0.000000
19	3ISE517, 3SFE517, 3IIS557, 3SFE541, 3SFE513, 3ISY557, 3ISE550, 3ISE541, 3ISE513, 3SFE550	10	2	3SFE517, [3ISE517]	2	1,2	1.000000	0.200000	1.000000
20	3ISE541, 3SFE541	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
21	3SFE540, 3SFE542, 3IIS558, 3ISE515, 3IIS553, 3IIS551, 3IIS503, 3ISE541, 3ISE514, 3ISE510, 3MTS590, 3ISY558, 3ISE517, 3SFE515, 3MTS595, 3ISY553, 3ISY551, 3MMC502, 3MTS581, 3SFE541, 3ECE515, 3SFE514,3ECE517, 3ECE514, 3ECE521, 3MMC503, 3SFE510, 3SFE517, 3SFE509, 3ISE550, 3SFE550	31	3	[3SFE541, 3ISE541, 3SFE542]	3	2,8,20	1.000000	0.096774	0.500000
					Year 1 -P Average:		**0.646032**	**0.424979**	**0.666667**
					MRR Without 'no answer' queries				0.736842

Study 2 SemaCS search (no personalisation): Year 1 Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec		1.0	1.0	1.0	1.0	1.0	0.9	0.8	0.8	0.8	0.8
Q 2 Prec				0.3		0.3			0.0		0.0
Q 3 Prec	0.0										
Q 4 Prec				1.0		1.0			0.0		0.0
Q 5 Prec				1.0				0.2			0.1
Q 6 Prec			0.5		0.0		0.0		0.0		0.0
Q 7 Prec						1.0					1.0
Q 8 Prec			0.5		0.0		0.0		0.0		0.0
Q 9 Prec						0.0					0.0
Q 10 Prec						0.3					0.3
Q 11 Prec						1.0					0.7
Q 12 Prec						1.0					1.0
Q 13 Prec				1.0		1.0			0.0		0.0
Q 14 Prec				1.0				1.0			0.0
Q 15 Prec	0.0										
Q 16 Prec						1.0					1.0
Q 17 Prec						1.0					1.0
Q 18 Prec						0.0					0.0
Q 19 Prec						1.0					1.0
Q 20 Prec						1.0					1.0
Q 21 Prec				0.5				0.3			0.2

Study 2 SemaCS search (no personalisation): Year 1 Interpolated average Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.8	0.8	0.8	0.8
Q 2 Prec	0.3	0.3	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0
Q 3 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 4 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 5 Prec	1.0	1.0	1.0	1.0	0.2	0.2	0.2	0.2	0.1	0.1	0.1
Q 6 Prec	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 7 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 8 Prec	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 9 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 10 Prec	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Q 11 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.7	0.7	0.7	0.7	0.7
Q 12 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 13 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 14 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0
Q 15 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 16 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 17 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 18 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 19 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 20 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 21 Prec	0.5	0.5	0.5	0.5	0.3	0.3	0.3	0.3	0.2	0.2	0.2
Average	0.67	0.67	0.67	0.62	0.57	0.57	0.44	0.43	0.38	0.38	0.38

Appendix B4 C: SemaCS (personalisation) Year 1 P/R and F-score

Q-ID	Student Query description	Student Query	Expert generated query matches (ID)	SemaCS generated (personalisation)	matched	m count	count	recall	precision	f-score
1	object oriented programming modules	oo programming	3ISY557, 3SFE513, 3SFE517, 3SFE541, 3SFE550, [3ISE550, 3ISE517, 3ISE541, 3IIS557, 3ISE513]	32:[3SFE541] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 3:[3SFE517, 3ISE517] 2:[3SFE504, 3ISE541] 1:[3IIS557, 3ISE513]	10	10	14	1	0.714285714	0.833333333
2	a module teaching rules and methods of graphical user interface designand implementation	interface design	3SFE509, [3SFE542, 3ISY557, 3IIS557]	34:[3SFE542] 32:[3SFE509, 3SFE541] 30:[3SFE510] 28:[3ISY557] 26:[3MTS580] 20:[3MMC501, 3SFE504] 18:[3ECE516] 17:[3ECE517] 16:[3SFE514, 3ISE514] 11:[3ECE515] 9:[3ISE550, 3ISY558, 3SFE550, 3ECE514] 8:[3SFE513, 3ISY599] 7:[3SFE530] 4:[3ISY509] 3:[3MTS590, 3MTS595]	3	4	23	0.75	0.130434783	0.222222222
3	animation	computing	none	10:[3SFE504] 8:[3SFE513]	0	0	2	0	0	0
4	graphical design	graphics	3SFE508, [3MTS580, 3ISY557, 3IIS557]	13:[3MTS580] 12:[3SFE508]	2	4	2	0.5	1	0.666666667
5	html java	web design	3SFE517, [3ISE517, 3MTS590]	32:[3SFE541] 17:[3ECE517, 3SFE542] 16:[3SFE509] 15:[3SFE510] 14:[3ISY557] 13:[3MTS580] 11:[3ECE515] 10:[3SFE504, 3MMC501] 9:[3ECE514, 3ECE516] 8:[3SFE514, 3ISE514] 6:[3MTS590] 3:[3MTS595]	1	3	16	0.333333333	0.0625	0.105263158
6	internet programming	java	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	14:[3ISY557] 4:[3SFE540]	2	5	2	0.4	1	0.571428571
7	internet programming	internet programming	3SFE517, [3ISE517]	32:[3SFE541] 17:[3MMC502] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 6:[3ISE517, 3SFE517] 2:[3ISE541] 1:[3IIS557, 3ISE513]	2	2	14	1	0.142857143	0.25
8	java programming	java	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	14:[3ISY557] 4:[3SFE540]	2	5	2	0.4	1	0.571428571
9	programming	c#	3SFE541, [3ISE541]	none	0	2	0	0	0	0

10	programming	c# programming	3SFE541, [3ISE541]	32:[3SFE541] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 3:[3SFE517, 3ISE517] 2:[3ISE541] 1:[3IIS557, 3ISE513]	2	2	13	1	0.153846154	0.266666667
11	i want to get somw inframtion on this module	Rapid Application Dev	3ISY555, [3IIS555]	64:[3SFE541] 30:[3SFE510] 20:[3SFE504] 17:[3SFE542] 16:[3SFE509] 14:[3ISY557] 13:[3MTS580] 12:[3SFE508] 11:[3ECE515] 10:[3MMC501] 9:[3ECE516, 3ISE550, 3ISY558, 3SFE550] 8:[3SFE513, 3SFE514, 3ISE514, 3ISY599] 7:[3SFE530] 5:[3IIS555] 4:[3SFE540, 3ISY509] 3:[3ISY555, 3SFE517, 3ISE517, 3ISE510]	2	2	26	1	0.076923077	0.142857143
12		Database systems	3ISY553, [3IIS553]	17:[3MMC502] 15:[3SFE510] 12:[3SFE508] 10:[3MMC501] 9:[3ISY558] 8:[3ECE523] 6:[3IIS553, 3ISY553] 4:[3MMC500] 2:[3ISY551, 3ISE518, 3IIS551, 3SFE518] 1:[3MTS572]	2	2	14	1	0.142857143	0.25
13	Interested in learning graphics	Graphics	3SFE508, [3MTS580, 3ISY557, 3IIS557]	13:[3MTS580] 12:[3SFE508]	2	4	2	0.5	1	0.666666667
14	php	internet	3SFE517, [3ISE517, 3MMC502]	17:[3MMC502] 3:[3ISE517, 3SFE517]	3	3	3	1	1	1
15	i want to make games	game	none	11:[3ECE515] 9:[3ECE514]	0	0	2	0	0	0
16	i want to make games	3d	3SFE508, 3MTS580	13:[3MTS580] 12:[3SFE508]	2	2	2	1	1	1
17	i would like to learn silerlight .	.net	3SFE541, [3ISE541]	32:[3SFE541] 2:[3ISE541]	2	2	2	1	1	1
18	programming in C++	c++	3SFE504, [3ISE504]	19:[3SFE504] 17:[3SFE541] 15:[3ISY509] 14:[3ISE515, 3ISE514] 13:[3ECE515] 12:[3SFE515, 3SFE514, 3ISY557, 3ECE517] 11:[3ECE514, 3ECE508, 3SFE542] 10:[3ISY599] 8:[3SFE530, 3SFE510, 3ECE516, 3ISE550] 7:[3ISY558, 3MTS580] 6:[3SFE550, 3ECE523, 3SFE509] 5:[3SFE513, 3MMC599, 3SFE508, 3ISY560, 3MMC502] 3:[3MMC501, 3IIS509, 3ISE630, 3ISE513, 3IIS552, 3IIS557] 1:[3MTS597, 3ISY552,	1	2	38	0.5	0.026315789	0.05

				3ECE521, 3MTS581]						
19	LEARN HOW TO CREATE WEBSITES USING CODE	INTERNET PROGRAMMING	3SFE517, [3ISE517]	32:[3SFE541] 17:[3MMC502] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 6:[3ISE517, 3SFE517] 2:[3ISE541] 1:[3IIS557, 3ISE513]	2	2	14	1	0.142857143	0.25
20	LEARN PROGRAMMING LANGUAGES	.NET	3SFE541, [3ISE541]	32:[3SFE541] 2:[3ISE541]	2	2	2	1	1	1
21	looking for my module (3sfe542)	mobile web xml xslt	[3SFE541, 3ISE541, 3SFE542]	96:[3SFE541] 17:[3SFE542] 14:[3ISY557] 10:[3MMC501] 8:[3SFE540] 3:[3MTS590]	2	3	6	0.666666667	0.333333333	0.444444444

Appendix B4 D: SemaCS (personalisation) Year 1 interpolated average P/R and MRR

Study 2 SemaCS search (with personalisation): Recall Precision and MRR									
Q №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/ MRR
1	32:[3SFE541] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 3:[3SFE517, 3ISE517] 2:[3SFE504, 3ISE541] 1:[3IIS557, 3ISE513]	14	10	3ISY557, 3SFE513, 3SFE517, 3SFE541, 3SFE550, [3ISE550, 3ISE517, 3ISE541, 3IIS557, 3ISE513]	10	1,2,5,6,7, 9,10,12,1 3,14	1.000000	0.714286	1.000000
2	34:[3SFE542] 32:[3SFE509, 3SFE541] 30:[3SFE510] 28:[3ISY557] 26:[3MTS580] 20:[3MMC501, 3SFE504] 18:[3ECE516] 17:[3ECE517] 16:[3SFE514, 3ISE514] 11:[3ECE515] 9:[3ISE550, 3ISY558, 3SFE550, 3ECE514] 8:[3SFE513, 3ISY599] 7:[3SFE530] 4:[3ISY509] 3:[3MTS590, 3MTS595]	23	4	3SFE509, [3SFE542, 3ISY557, 3IIS557]	3	1,2,5	0.750000	0.130435	1.000000
3	10:[3SFE504] 8:[3SFE513]	2	0	none	0	0	0.000000	0.000000	0.000000
4	13:[3MTS580] 12:[3SFE508]	2	4	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2	1,2	0.500000	1.000000	1.000000
5	32:[3SFE541] 17:[3ECE517, 3SFE542] 16:[3SFE509] 15:[3SFE510] 14:[3ISY557] 13:[3MTS580] 11:[3ECE515] 10:[3SFE504, 3MMC501] 9:[3ECE514, 3ECE516] 8:[3SFE514, 3ISE514] 6:[3MTS590] 3:[3MTS595]	16	3	3SFE517, [3ISE517, 3MTS590]	1	15	0.333333	0.062500	0.000000
6	14:[3ISY557] 4:[3SFE540]	2	5	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	2	1,2	0.400000	1.000000	1.000000
7	32:[3SFE541] 17:[3MMC502] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 6:[3ISE517, 3SFE517] 2:[3ISE541] 1:[3IIS557, 3ISE513]	14	2	3SFE517, [3ISE517]	2	10,11	1.000000	0.142857	0.000000
8	14:[3ISY557] 4:[3SFE540]	2	5	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	2	1,2	0.400000	1.000000	1.000000
9	none	0	2	3SFE541, [3ISE541]	0	0	0.000000	0.000000	0.000000
10	32:[3SFE541] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 3:[3SFE517, 3ISE517] 2:[3ISE541] 1:[3IIS557, 3ISE513]	13	2	3SFE541, [3ISE541]	2	1,11	1.000000	0.153846	1.000000

11	64:[3SFE541] 30:[3SFE510] 20:[3SFE504] 17:[3SFE542] 16:[3SFE509] 14:[3ISY557] 13:[3MTS580] 12:[3SFE508] 11:[3ECE515] 10:[3MMC501] 9:[3ECE516, 3ISE550, 3ISY558, 3SFE550] 8:[3SFE513, 3SFE514, 3ISE514, 3ISY599] 7:[3SFE530] 5:[3IIS555] 4:[3SFE540, 3ISY509] 3:[3ISY555, 3SFE517, 3ISE517, 3ISE510]	26	2	3ISY555, [3IIS555]	2	20,23	1.000000	0.076923	0.000000
12	17:[3MMC502] 15:[3SFE510] 12:[3SFE508] 10:[3MMC501] 9:[3ISY558] 8:[3ECE523] 6:[3IIS553, 3ISY553] 4:[3MMC500] 2:[3ISY551, 3ISE518, 3IIS551, 3SFE518] 1:[3MTS572]	14	2	3ISY553, [3IIS553]	2	7,8	1.000000	0.142857	0.000000
13	13:[3MTS580] 12:[3SFE508]	2	4	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2	1,2	0.500000	1.000000	1.000000
14	17:[3MMC502] 3:[3ISE517, 3SFE517]	3	3	3SFE517, [3ISE517, 3MMC502]	3	1,2,3	1.000000	1.000000	1.000000
15	11:[3ECE515] 9:[3ECE514]	2	0	none	0	0	0.000000	0.000000	0.000000
16	13:[3MTS580] 12:[3SFE508]	2	2	3SFE508, 3MTS580	2	1,2	1.000000	1.000000	1.000000
17	32:[3SFE541] 2:[3ISE541]	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
18	19:[3SFE504] 17:[3SFE541] 15:[3ISY509] 14:[3ISE515, 3ISE514] 13:[3ECE515] 12:[3SFE515, 3SFE514, 3ISY557, 3ECE517] 11:[3ECE514, 3ECE508, 3SFE542] 10:[3ISY599] 8:[3SFE530, 3SFE510, 3ECE516, 3ISE550] 7:[3ISY558, 3MTS580] 6:[3SFE550, 3ECE523, 3SFE509] 5:[3SFE513, 3MMC599, 3SFE508, 3ISY560, 3MMC502] 3:[3MMC501, 3IIS509, 3ISE630, 3ISE513, 3IIS552, 3IIS557] 1:[3MTS597, 3ISY552, 3ECE521, 3MTS581]	38	2	3SFE504, [3ISE504]	1	1	0.500000	0.026316	1.000000
19	32:[3SFE541] 17:[3MMC502] 14:[3ISY557] 13:[3MTS580] 9:[3ECE516, 3SFE550, 3ISE550] 8:[3SFE513, 3ECE523] 6:[3ISE517, 3SFE517] 2:[3ISE541] 1:[3IIS557, 3ISE513]	14	2	3SFE517, [3ISE517]	2	10,11	1.000000	0.142857	0.000000
20	32:[3SFE541] 2:[3ISE541]	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
21	96:[3SFE541] 17:[3SFE542] 14:[3ISY557] 10:[3MMC501] 8:[3SFE540] 3:[3MTS590]	6	3	[3SFE541, 3ISE541, 3SFE542]	2	1,2	0.666667	0.333333	1.000000
					Year 1 -P Average:		**0.669048**	**0.472677**	**0.619048**
					MRR Without 'no answer' queries				0.684211

Study 2 SemaCS search (with personalisation): Year 1 Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec		1.0	1.0	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7
Q 2 Prec				1.0		1.0			0.6		0.0
Q 3 Prec	0.0										
Q 4 Prec				1.0		1.0			0.0		0.0
Q 5 Prec				0.1				0.0			0.0
Q 6 Prec			1.0		1.0		0.0		0.0		0.0
Q 7 Prec						0.1					0.2
Q 8 Prec			1.0		1.0		0.0		0.0		0.0
Q 9 Prec						0.0					0.0
Q 10 Prec						1.0					0.2
Q 11 Prec						0.1					0.1
Q 12 Prec						0.1					0.3
Q 13 Prec				1.0		1.0			0.0		0.0
Q 14 Prec				1.0				1.0			1.0
Q 15 Prec	0.0										
Q 16 Prec						1.0					1.0
Q 17 Prec						1.0					1.0
Q 18 Prec						1.0					0.0
Q 19 Prec						0.1					0.2
Q 20 Prec						1.0					1.0
Q 21 Prec				1.0				1.0			0.0

Study 2 SemaCS search (with personalisation): Year 1 interpolated average Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec	1.0	1.0	1.0	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
Q 2 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.6	0.6	0.6	0.0	0.0
Q 3 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 4 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 5 Prec	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 6 Prec	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 7 Prec	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Q 8 Prec	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 9 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 10 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.2	0.2	0.2	0.2	0.2
Q 11 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Q 12 Prec	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Q 13 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 14 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 15 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 16 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 17 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 18 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 19 Prec	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Q 20 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 21 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0
Average	0.65	0.65	0.65	0.64	0.63	0.54	0.34	0.34	0.29	0.27	0.27

Appendix B4 E: SRS search Year 1 P/R and F-score

Q-ID	Student Query description	Student Query	Expert generated query matches (ID)	SRS Generated	matched	mcount	count	recall	precision	f-score
1	object oriented programming modules	oo programming	3ISY557, 3SFE513, 3SFE517, 3SFE541, 3SFE550, [3ISE550, 3ISE517, 3ISE541, 3IIS557, 3ISE513]	none	0	10	0	0	0	0
2	a module teaching rules and methods of graphical user interface designand implementation	interface design	3SFE509, [3SFE542, 3ISY557, 3IIS557]	3SFE509	1	4	1	0.25	1	0.4
3	animation	computing	none	none	0	0	0	0	0	0
4	graphical design	graphics	3SFE508, [3MTS580, 3ISY557, 3IIS557]	3MTS580, 3SFE508	2	4	2	0.5	1	0.666666667
5	html java	web design	3SFE517, [3ISE517, 3MTS590]	3MTS590	1	3	1	0.333333333	1	0.5
6	internet programming	java	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	3SFE540	1	5	1	0.2	1	0.333333333
7	internet programming	internet programming	3SFE517, [3ISE517]	3SFE517, 3ISE517	2	2	2	1	1	1
8	java programming	java	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	3SFE540	1	5	1	0.2	1	0.333333333
9	programming	c#	3SFE541, [3ISE541]	3SFE541, 3ISE541	2	2	2	1	1	1
10	programming	c# programming	3SFE541, [3ISE541]	none	0	2	0	0	0	0
11	i want to get somw inframtion on this module	Rapid Application Dev	3ISY555, [3IIS555]	3ISY555, 3IIS555	2	2	2	1	1	1
12		Database systems	3ISY553, [3IIS553]	none	0	2	0	0	0	0
13	Interested in learning graphics	Graphics	3SFE508, [3MTS580, 3ISY557, 3IIS557]	3SFE508, 3MTS580	2	4	2	0.5	1	0.666666667
14	php	internet	3SFE517, [3ISE517, 3MMC502]	3SFE517, 3ISE517	2	3	2	0.666666667	1	0.8
15	i want to make games	game	none	none	0	0	0	0	0	0
16	i want to make games	3d	3SFE508, 3MTS580	3MTS580	1	2	1	0.5	1	0.666666667

17	i would like to learn silerlight .	.net	3SFE541, [3ISE541]	3SFE541, 3ISE541	2	2	2	1	1	1
18	programming in C++	c++	3SFE504, [3ISE504]	none	0	2	0	0	0	0
19	LEARN HOW TO CREATE WEBSITES USING CODE	INTERNET PROGRAMMING	3SFE517, [3ISE517]	3SFE517, 3ISE517	2	2	2	1	1	1
20	LEARN PROGRAMMING LANGUAGES	.NET	3SFE541, [3ISE541]	3SFE541, 3ISE541	2	2	2	1	1	1
21	looking for my module (3sfe542)	mobile web xml xslt	[3SFE541, 3ISE541, 3SFE542]	none	0	3	0	0	0	0

Appendix B4 F: SRS search Year 1 interpolated average P/R and MRR

Study 2 UoW SRS search: Recall Precision and MRR									
Q №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/ MRR
1	none	0	10	3ISY557, 3SFE513, 3SFE517, 3SFE541, 3SFE550, [3ISE550, 3ISE517, 3ISE541, 3IIS557, 3ISE513]	0	0	0.000000	0.000000	0.000000
2	3SFE509	1	4	3SFE509, [3SFE542, 3ISY557, 3IIS557]	1	1	0.250000	1.000000	1.000000
3	none	0	0	none	0	0	0.000000	0.000000	0.000000
4	3MTS580, 3SFE508	2	4	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2	1,2	0.500000	1.000000	1.000000
5	3MTS590	1	3	3SFE517, [3ISE517, 3MTS590]	1	1	0.333333	1.000000	1.000000
6	3SFE540	1	5	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	1	1	0.200000	1.000000	1.000000
7	3SFE517, 3ISE517	2	2	3SFE517, [3ISE517]	2	1,2	1.000000	1.000000	1.000000
8	3SFE540	1	5	3SFE550, 3SFE540, [3ISE550, 3ISY557, 3IIS557]	1	1	0.200000	1.000000	1.000000
9	3SFE541, 3ISE541	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
10	none	0	2	3SFE541, [3ISE541]	0	0	0.000000	0.000000	0.000000
11	3ISY555, 3IIS555	2	2	3ISY555, [3IIS555]	2	1,2	1.000000	1.000000	1.000000
12	none	0	2	3ISY553, [3IIS553]	0	0	0.000000	0.000000	0.000000
13	3SFE508, 3MTS580	2	4	3SFE508, [3MTS580, 3ISY557, 3IIS557]	2	1,2	0.500000	1.000000	1.000000
14	3SFE517, 3ISE517	2	3	3SFE517, [3ISE517, 3MMC502]	2	1,2	0.666667	1.000000	1.000000
15	none	0	0	none	0	0	0.000000	0.000000	0.000000
16	3MTS580	1	2	3SFE508, 3MTS580	1	1	0.500000	1.000000	1.000000
17	3SFE541, 3ISE541	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
18	none	0	2	3SFE504, [3ISE504]	0	0	0.000000	0.000000	0.000000
19	3SFE517, 3ISE517	2	2	3SFE517, [3ISE517]	2	1,2	1.000000	1.000000	1.000000
20	3SFE541, 3ISE541	2	2	3SFE541, [3ISE541]	2	1,2	1.000000	1.000000	1.000000
21	none	0	3	[3SFE541, 3ISE541, 3SFE542]	0	0	0.000000	0.000000	0.000000
					Year 1 -P Average:		0.435714	0.666667	0.666667
					MRR Without 'no answer' queries				0.736842

Study 2 UoW SRS search: Year 1 Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 2 Prec				1.0		0.0			0.0		0.0
Q 3 Prec	0.0										
Q 4 Prec				1.0		1.0			0.0		0.0
Q 5 Prec				1.0				0.0			0.0
Q 6 Prec			1.0		0.0		0.0		0.0		0.0
Q 7 Prec						1.0					1.0
Q 8 Prec			1.0		0.0		0.0		0.0		0.0
Q 9 Prec						1.0					1.0
Q 10 Prec						0.0					0.0
Q 11 Prec						1.0					1.0
Q 12 Prec						0.0					0.0
Q 13 Prec				1.0		1.0			0.0		0.0
Q 14 Prec				1.0				1.0			0.0
Q 15 Prec	0.0										
Q 16 Prec						1.0					0.0
Q 17 Prec						1.0					1.0
Q 18 Prec						0.0					0.0
Q 19 Prec						1.0					1.0
Q 20 Prec						1.0					1.0
Q 21 Prec				0.0				0.0			0.0

Study 2 UoW SRS search: Year 1 Interpolated average Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 1 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 2 Prec	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 3 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 4 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 5 Prec	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 6 Prec	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 7 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 8 Prec	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 9 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 10 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 11 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 12 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 13 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 14 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0
Q 15 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 16 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 17 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 18 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 19 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 20 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 21 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	0.67	0.67	0.67	0.57	0.48	0.48	0.33	0.33	0.29	0.29	0.29

Appendix B5: University of Westminster SRS case study Year 2 log analyses

University of Westminster SRS study Year 2 interpolated average Precision Recall											
Recall-->	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
SemaCS no personalisation	0.3545	0.3545	0.3545	0.3545	0.3110	0.3110	0.1367	0.0932	0.0466	0.0466	0.0466
SemaCS with personalisation	0.5196	0.5196	0.5109	0.5109	0.4964	0.4964	0.4564	0.3781	0.3817	0.3382	0.2948
UoW SRS search	0.2609	0.2609	0.2609	0.2609	0.2174	0.2174	0.1304	0.0435	0.0000	0.0000	0.0000

University of Westminster SRS Year 2 average				
	Recall	**Precision**	**MRR**	**F-score**
SemaCS no personalisation	0.347308	0.309300	0.340580	0.327204
SemaCS with personalisation	0.662267	0.371536	0.521739	0.476021
UoW SRS search	0.135352	0.260870	0.260870	0.178230

Appendix B5 A: SemaCS (no personalisation) Year 2 P/R and F-score

Q-ID	Student Query description	Student Query	Expert generated query matches (ID)	SemaCS no personalisation	Matched #	Res Count	Match #	Recall	Precision	F-Score
22	applications	internet	3SFE617, [3ISE617, 3SFE615]	2:[3ISE617] 1:[3SFE617]	2	2	3	0.6666667	1	0.8
23	programming language	ada	none	2:[3SFE618]	0	1	0	0	0	0
24	programming language	occam	none	2:[3MTS621] 1:[3MTS699, 3ECE603]	0	3	0	0	0	0
25		business	3ISY612, [3IIS612, 3IIS655, 3ISY655, 3ECE615, 3ISY657, 3IIS657]	4:[3IIS655] 3:[3IIS612] 2:[3ISY655] 1:[3ISY612]	4	4	7	0.5714286	1	0.7272727
26	I would like to know more about web design	Web Design	3SFE617, [3ISE617, 3SFE620]	6:[3MTS694, 3ISE611, 3ISE601, 3SFE611, 3SFE602, 3ECE624, 3SFE601, 3ECE621] 4:[3IIS657, 3ISY657, 3ISE617, 3SFE617, 3ECE602, 3MTS697] 3:[3ISE699, 3SFE699, 3SMT602, 3SFE621] 2:[3SFE615, 3ISY608, 3ISY651, 3IIS652, 3IIS651, 3ISY652, 3ISY699, 3MTS622, 3SFE619, 3ISE619, 3ECE616, 3MTS695, 3EDM671, 3MMC606, 3ECM100] 1:[3SFE620, 3IIS654, 3ISY654, 3TSE613, 3SFE630, 3MTS696, 3IIS699, 3EDM672]	3	#	3	1	0.0731707	0.1363636
27	i would like to know more about php and perl	internet programming	3SFE617, [3ISE617]	3:[3ISE617] 2:[3MTS621, 3SFE605, 3SFE610] 1:[3SFE617, 3MTS690]	2	6	2	1	0.3333333	0.5
28		graphics	3MTS626, [3MTS691]	1:[3MTS626]	1	1	2	0.5	1	0.6666667
29	web designing	web designing	3SFE617, [3ISE617, 3SFE620]	4:[3IIS657, 3ISY657, 3MTS697] 3:[3MTS694, 3ISE611, 3ISE601, 3SFE611, 3SFE602, 3ECE624, 3SFE601, 3ECE621, 3ISE699, 3SFE699, 3SMT602, 3SFE621] 2:[3SFE615, 3ISY608, 3ISY651, 3IIS652, 3IIS651, 3ISY652, 3ISY699, 3MTS622, 3ISE617, 3SFE617, 3ECE602, 3SFE619, 3ISE619, 3ECE616, 3MTS695, 3EDM671, 3MMC606, 3ECM100, 3MTS674] 1:[3SFE620, 3IIS654, 3ISY654, 3TSE613, 3SFE630, 3MTS696, 3IIS699, 3EDM672]	3	#	3	1	0.0714286	0.1333333
30		website administration	3SFE615	2:[3MTS622, 3SFE615]	1	2	1	1	0.5	0.6666667

31	i would like to learn more about flash	learn flash	none [3MTS691]	3:[3MTS694] 1:[3TSE613]	0	2	1	0	0	0
32		programming	3SFE605, 3SFE610, [3MTS621, 3SFE621, 3SFE602, 3ECE624]	2:[3MTS621, 3SFE605, 3SFE610] 1:[3MTS690]	3	4	6	0.5	0.75	0.6
33	even though it is boring!	networks	3SFE611, [3ISE611, 3MMC601, 3SMC602, 3SFE622, 3TSE613, 3ISE619, 3SFE619]	2:[3MMC601] 1:[3SMC602]	2	2	8	0.25	1	0.4
34	a platform games	java games	none [3MTS626, 3MTS690]	3:[3ISE611, 3ISE699, 3SFE611, 3SFE699] 2:[3ISE616, 3SFE616] 1:[3IIS612, 3ISY612, 3IIS654, 3ISY654]	0	#	2	0	0	0
35	i would like to learn how to create a professional looking web site	html	none	3:[3MTS694]	0	1	0	0	0	0
36	javascript programme	java	3SFE605, [3SFE617, 3ISE617, 3MTS626]	2:[3ISE616, 3SFE616] 1:[3IIS612, 3ISY612]	0	4	4	0	0	0
37	help on actionscript	flash design	none [3MTS691]	3:[3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621] 2:[3ECE602, 3ISE617, 3SFE617] 1:[3TSE613]	0	#	1	0	0	0
38	i like to learn more about netwoks	network	3SFE611, [3ISE611, 3ISE619, 3MMC601, 3SFE619, 3SFE622, 3SMC602, 3TSE613]	3:[3ISE611, 3SFE611] 2:[3ISE619, 3SFE619]	4	4	8	0.5	1	0.6666667
39	graphic design	graphic design	none	3:[3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621] 2:[3MTS621, 3SFE605, 3SFE610, 3ECE602, 3ISE617, 3SFE617] 1:[3MTS690]	0	#	0	0	0	0

40	I would like to learn more about database and its design	database design	none	3:[3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621] 2:[3ISE616, 3SFE616, 3ECE602, 3ISE617, 3SFE617] 1:[3IIS612, 3ISY612]	0	#	0	0	0	0
41	learn more about 3d design	3d design	3MTS626, [3MTS691]	3:[3SMT602, 3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621] 2:[3SFE610, 3SFE605, 3MTS621, 3ECE603, 3MMC601, 3SFE609, 3ISE609, 3SRT601, 3CCE632, 3ECE602, 3ISE617, 3SFE617] 1:[3MTS690, 3SFE613, 3ISE613, 3SMC602]	0	#	2	0	0	0
42	I'm interested in databases	database systems	none	4:[3ISY657, 3IIS657] 3:[3SFE601, 3ISE601] 2:[3ISE616, 3SFE616, 3CCE632, 3SRT601, 3SFE622, 3SFE609, 3ISY699, 3ISY655, 3ISY652, 3ISY651, 3MTS629, 3ISE609, 3IIS655, 3IIS652, 3IIS651] 1:[3IIS612, 3ISY612, 3SMC601, 3TSE612]	0	#	0	0	0	0
43	3d, animation	computer graphics	3MTS626 [3MTS691]	3:[3SFE623, 3SFE630] 1:[3MTS626]	1	3	2	0.5	0.3333333	0.4
44	designing games with multiplayer feature	graphics 3d multi	3MTS626 [3MTS691]	3:[3SMT602] 2:[3SFE610, 3SFE605, 3MTS621, 3ECE603, 3MMC601, 3SFE609, 3ISE609, 3SRT601, 3CCE632, 3ISE616, 3SFE616] 1:[3MTS626, 3MTS690, 3SFE613, 3ISE613, 3SMC602, 3IIS612, 3ISY612]	1	#	2	0.5	0.0526316	0.0952381

Appendix B5 B: SemaCS (no personalisation) Year 2 interpolated average P/R and MRR

Study 2 SemaCS search (no personalisation): Year 2 Recall Precision and MRR									
Q №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
22	3ISE617, 3SFE617	2	3	3SFE617, [3ISE617, 3SFE615]	2	1,2	0.666667	1.000000	1.000000
23	3SFE618	1	0	none	0	0	0.000000	0.000000	0.000000
24	3MTS621, 3MTS699, 3ECE603	3	0	none	0	0	0.000000	0.000000	0.000000
25	3IIS655, 3IIS612, 3ISY655, 3ISY612	4	7	3ISY612, [3IIS612, 3IIS655, 3ISY655, 3ECE615, 3ISY657, 3IIS657]	4	1,2,3,4	0.571429	1.000000	1.000000
26	3MTS694, 3ISE611, 3ISE601, 3SFE611, 3SFE602, 3ECE624, 3SFE601, 3ECE621, 3IIS657, 3ISY657, 3ISE617, 3SFE617, 3ECE602, 3MTS697, 3ISE699, 3SFE699, 3SMT602, 3SFE621, 3SFE615, 3ISY608, 3ISY651, 3IIS652, 3IIS651, 3ISY652, 3ISY699, 3MTS622, 3SFE619, 3ISE619, 3ECE616, 3MTS695, 3EDM671, 3MMC606, 3ECM100, 3SFE620, 3IIS654, 3ISY654, 3TSE613, 3SFE630, 3MTS696, 3IIS699, 3EDM672	41	3	3SFE617, [3ISE617, 3SFE620]	3	11,12,34	1.000000	0.073171	0.000000
27	3ISE617, 3MTS621, 3SFE605, 3SFE610, 3SFE617, 3MTS690	6	2	3SFE617, [3ISE617]	2	1,5	1.000000	0.333333	1.000000
28	3MTS626	1	2	3MTS626, [3MTS691]	1	1	0.500000	1.000000	1.000000
29	3IIS657, 3ISY657, 3MTS697, 3MTS694, 3ISE611, 3ISE601, 3SFE611, 3SFE602, 3ECE624, 3SFE601, 3ECE621, 3ISE699, 3SFE699, 3SMT602, 3SFE621, 3SFE615, 3ISY608, 3ISY651, 3IIS652, 3IIS651, 3ISY652, 3ISY699, 3MTS622, 3ISE617, 3SFE617, 3ECE602, 3SFE619, 3ISE619, 3ECE616, 3MTS695, 3EDM671, 3MMC606, 3ECM100, 3MTS674, 3SFE620, 3IIS654, 3ISY654, 3TSE613, 3SFE630, 3MTS696, 3IIS699, 3EDM672	42	3	3SFE617, [3ISE617, 3SFE620]	3	25,26,36	1.000000	0.071429	0.000000
30	3MTS622, 3SFE615	2	1	3SFE615	1	2	1.000000	0.500000	0.500000
31	3MTS694, 3TSE613	2	1	none [3MTS691]	0	0	0.000000	0.000000	0.000000
32	3MTS621, 3SFE605, 3SFE610, 3MTS690	4	6	3SFE605, 3SFE610, [3MTS621, 3SFE621, 3SFE602, 3ECE624]	3	1,2,3	0.500000	0.750000	1.000000
33	3MMC601, 3SMC602	2	8	3SFE611, [3ISE611, 3MMC601, 3SMC602, 3SFE622, 3TSE613, 3ISE619, 3SFE619]	2	1,2	0.250000	1.000000	1.000000

34	3ISE611, 3ISE699, 3SFE611, 3SFE699, 3ISE616, 3SFE616, 3IIS612, 3ISY612, 3IIS654, 3ISY654	10	2	none [3MTS626, 3MTS690]	0	0	0.000000	0.000000	0.000000
35	3MTS694	1	0	none	0	0	0.000000	0.000000	0.000000
36	3ISE616, 3SFE616, 3IIS612, 3ISY612	4	4	3SFE605, [3SFE617, 3ISE617, 3MTS626]	0	0	0.000000	0.000000	0.000000
37	3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621, 3ECE602, 3ISE617, 3SFE617, 3TSE613	12	1	none [3MTS691]	0	0	0.000000	0.000000	0.000000
38	3ISE611, 3SFE611, 3ISE619, 3SFE619	4	8	3SFE611, [3ISE611, 3ISE619, 3MMC601, 3SFE619, 3SFE622, 3SMC602, 3TSE613]	4	1,2,3,4	0.500000	1.000000	1.000000
39	3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621, 3MTS621, 3SFE605, 3SFE610, 3ECE602, 3ISE617, 3SFE617, 3MTS690	15	0	none	0	0	0.000000	0.000000	0.000000
40	3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621, 3ISE616, 3SFE616, 3ECE602, 3ISE617, 3SFE617, 3IIS612, 3ISY612	15	0	none	0	0	0.000000	0.000000	0.000000
41	3SMT602, 3SFE611, 3SFE602, 3SFE601, 3MTS694, 3ISE611, 3ISE601, 3ECE624, 3ECE621, 3SFE610, 3SFE605, 3MTS621, 3ECE603, 3MMC601, 3SFE609, 3ISE609, 3SRT601, 3CCE632, 3ECE602, 3ISE617, 3SFE617, 3MTS690, 3SFE613, 3ISE613, 3SMC602	25	2	3MTS626, [3MTS691]	0	0	0.000000	0.000000	0.000000
42	3ISY657, 3IIS657, 3SFE601, 3ISE601, 3ISE616, 3SFE616, 3CCE632, 3SRT601, 3SFE622, 3SFE609, 3ISY699, 3ISY655, 3ISY652, 3ISY651, 3MTS629, 3ISE609, 3IIS655, 3IIS652, 3IIS651, 3IIS612, 3ISY612, 3SMC601, 3TSE612	23	0	none	0	0	0.000000	0.000000	0.000000
43	3SFE623, 3SFE630, 3MTS626	3	2	3MTS626 [3MTS691]	1	3	0.500000	0.333333	0.333333
44	3SMT602, 3SFE610, 3SFE605, 3MTS621, 3ECE603, 3MMC601, 3SFE609, 3ISE609, 3SRT601, 3CCE632, 3ISE616, 3SFE616, 3MTS626, 3MTS690, 3SFE613, 3ISE613, 3SMC602, 3IIS612, 3ISY612	19	2	3MTS626 [3MTS691]	1	13	0.500000	0.052632	0.000000
				Average:			**0.347308**	**0.309300**	**0.340580**
				MRR Without 'no answer' queries					0.460784

Study 2 SemaCS search (no personalisation): Year 2 Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 22 Prec				1.0				1.0			0.0
Q 23 Prec	0.0										
Q 24 Prec	0.0										
Q 25 Prec		1.0		1.0	1.0		1.0	0.0		0.0	0.0
Q 26 Prec				0.1				0.2			0.1
Q 27 Prec						1.0					0.4
Q 28 Prec						1.0					0.0
Q 29 Prec				0.0				0.1			0.1
Q 30 Prec											0.5
Q 31 Prec											0.0
Q 32 Prec			1.0	1.0		1.0		0.0	0.0		0.0
Q 33 Prec		1.0		1.0	0.0	0.0	0.0		0.0	0.0	0.0
Q 34 Prec						0.0					0.0
Q 35 Prec	0.0										
Q 36 Prec				0.0		0.0			0.0		0.0
Q 37 Prec											0.0
Q 38 Prec		1.0		1.0	1.0	1.0	0.0		0.0	0.0	0.0
Q 39 Prec	0.0										
Q 40 Prec	0.0										
Q 41 Prec						0.0					0.0
Q 42 Prec	0.0										
Q 43 Prec						0.3					0.0
Q 44 Prec						0.1					0.0

Study 2 SemaCS search (no personalisation): Year 2 interpolated average Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 22 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0
Q 23 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 24 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 25 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
Q 26 Prec	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1
Q 27 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.4	0.4	0.4	0.4	0.4
Q 28 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 29 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Q 30 Prec	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Q 31 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 32 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 33 Prec	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 34 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 35 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 36 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 37 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 38 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 39 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 40 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 41 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 42 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 43 Prec	0.3	0.3	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0
Q 44 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0
Average	0.35	0.35	0.35	0.35	0.31	0.31	0.14	0.09	0.05	0.05	0.05

Appendix B5 C: SemaCS (personalisation) Year 2 P/R and F-score

Q ID	Student Query description	Student Query	Expert generated query matches	SemaCS personalisation	Matched #	Res Count	Match #	Recall	Precision	F-Score
22	applications	internet	3SFE617, [3ISE617, 3SFE615]	5:[3ISE617, 3SFE615] 3:[3SFE617]	3	3	3	1	1	1
23	programming language	ada	none	33:[3SFE620]	0	1	0	0	0	0
24	programming language	occam	none	33:[3SFE620]	0	1	0	0	0	0
25		business	3ISY612, [3IIS612, 3IIS655, 3ISY655, 3ECE615, 3ISY657, 3IIS657]	10:[3ECE615] 7:[3ISY655] 6:[3IIS612, 3ISY657] 4:[3IIS655, 3ISY612]	6	6	7	0.8571429	1	0.9230769
26	I would like to know more about web design	Web Design	3SFE617, [3ISE617, 3SFE620]	66:[3SFE620] 38:[3ISE617, 3SFE617] 31:[3MTS674] 22:[3SFE602] 21:[3SFE618, 3ECE602] 20:[3ECE624, 3ECM100] 18:[3EEE612] 15:[3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601] 14:[3ECE621, 3CCE632] 13:[3SFE615, 3MTS696, 3ECE616] 11:[3MTS690, 3ISY651] 10:[3ECE603] 9:[3ISY612, 3IIS612] 2:[3MTS694, 3ISE611]	3	#	3	1	0.1111111	0.2
27	i would like to know more about php and perl	internet programming	3SFE617, [3ISE617]	28:[3SFE621] 22:[3SFE602] 20:[3ECE624] 18:[3SFE610] 17:[3SFE605, 3MTS621] 16:[3MTS691] 5:[3ISE617, 3SFE615] 3:[3SFE617]	2	#	2	1	0.2	0.3333333
28		graphics	3MTS626, [3MTS691]	13:[3MTS691] 2:[3MTS626]	2	2	2	1	1	1
29	web designing	web designing	3SFE617, [3ISE617, 3SFE620]	33:[3SFE620] 31:[3MTS674] 19:[3ISE617, 3SFE617] 17:[3ISY657] 13:[3SFE615]	3	6	3	1	0.5	0.6666667
30		website administration	3SFE615	8:[3EDM671] 5:[3ECE615, 3SFE613, 3MTS699, 3MTS690, 3ISE613, 3SFE615] 4:[3ISY652] 3:[3SFE630, 3SFE623, 3MTS696, 3EDM672] 2:[3ISE699, 3SFE699] 1:[3ISY699, 3SFE601, 3ISE601]	1	#	1	1	0.0588235	0.1111111
31	i would like to learn more about flash	learn flash	none [3MTS691]	13:[3MTS691] 10:[3SFE621] 5:[3MTS690] 4:[3ISY652]	1	4	1	1	0.25	0.4
32		programming	3SFE605, 3SFE610, [3MTS621, 3SFE621, 3SFE602, 3ECE624]	28:[3SFE621] 22:[3SFE602] 20:[3ECE624] 18:[3SFE610] 17:[3SFE605, 3MTS621] 16:[3MTS691]	6	7	6	1	0.8571429	0.9230769

33	even though it is boring!	networks	3SFE611, [3ISE611, 3MMC601, 3SMC602, 3SFE622, 3TSE613, 3ISE619, 3SFE619]	23:[3SFE622] 11:[3MTS623] 9:[3MMC601, 3TSE613] 7:[3SMC602]	4	5	8	0.5	0.8	0.6153846
34	a platform games	java games	none [3MTS626, 3MTS690]	19:[3ISE617, 3SFE617] 18:[3SFE622] 10:[3TSE613] 8:[3SMC601] 6:[3MTS626] 4:[3MTS690]	2	7	2	1	0.2857143	0.4444444
35	i would like to learn how to create a professional looking web site	html	none	15:[3ISE601, 3SFE601, 3SFE611]	0	3	0	0	0	0
36	javascript programme	java	3SFE605, [3SFE617, 3ISE617, 3MTS626]	19:[3ISE617, 3SFE617] 6:[3MTS626]	3	3	4	0.75	1	0.8571429
37	help on actionscript	flash design	none [3MTS691]	33:[3SFE620] 31:[3MTS674] 22:[3SFE602] 21:[3SFE618, 3ECE602] 20:[3ECE624, 3ECM100] 19:[3SFE617, 3ISE617] 18:[3EEE612] 16:[3MTS690] 15:[3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601] 14:[3ECE621, 3CCE632] 13:[3MTS691, 3MTS696, 3ECE616] 11:[3ISY651] 10:[3ECE603] 9:[3ISY612, 3IIS612] 4:[3ISY652] 2:[3MTS694, 3ISE611]	1	#	1	1	0.0357143	0.0689655
38	i like to learn more about netwoks	network	3SFE611, [3ISE611, 3ISE619, 3MMC601, 3SFE619, 3SFE622, 3SMC602, 3TSE613]	18:[3SFE622] 7:[3SFE619] 6:[3SFE611] 3:[3ISE611, 3ISE619]	5	5	8	0.625	1	0.7692308
39	graphic design	graphic design	none	33:[3SFE620] 31:[3MTS674] 29:[3ECM100] 22:[3SFE602] 21:[3SFE618, 3ECE602] 20:[3ECE624] 19:[3ECE616, 3SFE617, 3ISE617] 18:[3EEE612] 15:[3SFE622, 3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601] 14:[3ECE621, 3CCE632] 13:[3MTS696] 11:[3MTS690, 3ISY651] 10:[3ECE603] 9:[3ISY612, 3IIS612] 8:[3EDM671] 4:[3MMC606] 2:[3MTS694, 3ISE611]	0	#	0	0	0	0

40	I would like to learn more about database and its design	database design	none	33:[3SFE620] 31:[3MTS674] 22:[3SFE602] 21:[3SFE618, 3ECE602] 20:[3ECE624, 3ECM100] 19:[3SFE617, 3ISE617] 18:[3EEE612] 15:[3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601] 14:[3ECE621, 3CCE632] 13:[3SFE615, 3MTS696, 3ECE616] 11:[3MTS690, 3ISY651] 10:[3ECE603] 9:[3ISY612, 3IIS612] 2:[3MTS694, 3ISE611]	0	#	0	0	0	0
41	learn more about 3d design	3d design	3MTS626, [3MTS691]	44:[3SFE602] 40:[3ECE624] 33:[3SFE620] 31:[3MTS674] 30:[3MMC600] 28:[3SFE621, 3CCE632] 21:[3SFE618, 3ECE602] 20:[3ECM100] 19:[3SFE617, 3ISE617] 18:[3SFE610, 3EEE612] 17:[3SFE605, 3MTS621] 16:[3MTS691, 3SRT601] 15:[3SFE601, 3SFE611, 3CCE633, 3ISE601] 14:[3ECE621] 13:[3MTS696, 3ECE616] 12:[3SFE616, 3SFE609, 3ISE609] 11:[3MTS690, 3ISY651] 10:[3ECE603] 9:[3ISY612, 3IIS612] 6:[3MTS626] 2:[3MTS694, 3ISE611]	1	#	2	0.5	0.0277778	0.0526316
42	I'm interested in databases	database systems	none	23:[3SFE622] 21:[3ECE602] 17:[3ISY657] 16:[3SRT601] 15:[3ISE601, 3SFE601, 3MMC600] 14:[3ECE621, 3CCE632, 3MTS629] 13:[3SFE615, 3ISY654] 12:[3SFE609, 3ECE615, 3ISE609] 11:[3ISY655, 3ISY651, 3SFE623] 10:[3ISY608] 9:[3SMT601, 3TSE613, 3TSE612, 3ISY612, 3IIS612] 7:[3ISY652, 3SMC601] 5:[3ISY699] 3:[3IIS657] 2:[3IIS655] 1:[3IIS652, 3IIS651]	0	#	0	0	0	0
43	3d, animation	computer graphics	3MTS626 [3MTS691]	26:[3MTS691] 11:[3SFE623, 3SRT601] 8:[3SFE630] 5:[3MTS629] 3:[3ECE624] 2:[3MTS626]	2	7	2	1	0.2857143	0.4444444
44	designing games with multiplayer feature	graphics 3d multi	3MTS626 [3MTS691]	42:[3MTS691] 28:[3SFE621] 22:[3SFE602] 20:[3ECE624] 18:[3SFE610] 17:[3SFE605, 3MTS621] 16:[3SRT601] 15:[3MMC600] 14:[3CCE632] 12:[3SFE616, 3SFE609, 3ISE609] 10:[3MTS674] 8:[3MTS626]	2	#	2	1	0.1333333	0.2352941

Appendix B5 D: SemaCS (personalisation) Year 2 interpolated average P/R and MRR

Study 2 SemaCS search (with personalisation): Year 2 Recall Precision and MRR									
Q №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
22	3ISE617, 3SFE615, 3SFE617	3	3	3SFE617, [3ISE617, 3SFE615]	3	1,2,3	1.000000	1.000000	1.000000
23	3SFE620	1	0	none	0	0	0.000000	0.000000	0.000000
24	3SFE620	1	0	none	0	0	0.000000	0.000000	0.000000
25	3ECE615, 3ISY655, 3IIS612, 3ISY657, 3IIS655, 3ISY612	6	7	3ISY612, [3IIS612, 3IIS655, 3ISY655, 3ECE615, 3ISY657, 3IIS657]	6	1,2,3,4,5, 6	0.857143	1.000000	1.000000
26	3SFE620, 3ISE617, 3SFE617, 3MTS674, 3SFE602, 3SFE618, 3ECE602, 3ECE624, 3ECM100, 3EEE612, 3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601, 3ECE621, 3CCE632, 3SFE615, 3MTS696, 3ECE616, 3MTS690, 3ISY651, 3ECE603, 3ISY612, 3IIS612, 3MTS694, 3ISE611	27	3	3SFE617, [3ISE617, 3SFE620]	3	1,2,3	1.000000	0.111111	1.000000
27	3SFE621, 3SFE602, 3ECE624, 3SFE610, 3SFE605, 3MTS621, 3MTS691, 3ISE617, 3SFE615, 3SFE617	10	2	3SFE617, [3ISE617]	2	8,10	1.000000	0.200000	0.000000
28	3MTS691, 3MTS626	2	2	3MTS626, [3MTS691]	2	1,2	1.000000	1.000000	1.000000
29	3SFE620, 3MTS674, 3ISE617, 3SFE617, 3ISY657, 3SFE615	6	3	3SFE617, [3ISE617, 3SFE620]	3	1,3,4	1.000000	0.500000	1.000000
30	3EDM671, 3ECE615, 3SFE613, 3MTS699, 3MTS690, 3ISE613, 3SFE615, 3ISY652, 3SFE630, 3SFE623, 3MTS696, 3EDM672, 3ISE699, 3SFE699, 3ISY699, 3SFE601, 3ISE601	17	1	3SFE615	1	7	1.000000	0.058824	0.000000
31	3MTS691, 3SFE621, 3MTS690, 3ISY652	4	1	none [3MTS691]	1	1	1.000000	0.250000	1.000000
32	3SFE621, 3SFE602, 3ECE624, 3SFE610, 3SFE605, 3MTS621, 3MTS691	7	6	3SFE605, 3SFE610, [3MTS621, 3SFE621, 3SFE602, 3ECE624]	6	1,2,3,4,5, 6	1.000000	0.857143	1.000000
33	3SFE622, 3MTS623, 3MMC601, 3TSE613, 3SMC602	5	8	3SFE611, [3ISE611, 3MMC601, 3SMC602, 3SFE622, 3TSE613, 3ISE619, 3SFE619]	4	1,3,4,5	0.500000	0.800000	1.000000
34	3ISE617, 3SFE617, 3SFE622, 3TSE613, 3SMC601, 3MTS626, 3MTS690	7	2	none [3MTS626, 3MTS690]	2	6,7	1.000000	0.285714	0.000000
35	3ISE601, 3SFE601, 3SFE611	3	0	none	0	0	0.000000	0.000000	0.000000
36	3ISE617, 3SFE617, 3MTS626	3	4	3SFE605, [3SFE617, 3ISE617, 3MTS626]	3	1,2,3	0.750000	1.000000	1.000000

37	3SFE620, 3MTS674, 3SFE602, 3SFE618, 3ECE602, 3ECE624, 3ECM100, 3SFE617, 3ISE617, 3EEE612, 3MTS690, 3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601, 3ECE621, 3CCE632, 3MTS691, 3MTS696, 3ECE616, 3ISY651, 3ECE603, 3ISY612, 3IIS612, 3ISY652, 3MTS694, 3ISE611	28	1	none [3MTS691]	1	19	1.000000	0.035714	0.000000
38	3SFE622, 3SFE619, 3SFE611, 3ISE611, 3ISE619	5	8	3SFE611, [3ISE611, 3ISE619, 3MMC601, 3SFE619, 3SFE622, 3SMC602, 3TSE613]	5	1,2,3,4,5	0.625000	1.000000	1.000000
39	3SFE620, 3MTS674, 3ECM100, 3SFE602, 3SFE618, 3ECE602, 3ECE624, 3ECE616, 3SFE617, 3ISE617, 3EEE612, 3SFE622, 3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601, 3ECE621, 3CCE632, 3MTS696, 3MTS690, 3ISY651, 3ECE603, 3ISY612, 3IIS612, 3EDM671, 3MMC606, 3MTS694, 3ISE611	29	0	none	0	0	0.000000	0.000000	0.000000
40	3SFE620, 3MTS674, 3SFE602, 3SFE618, 3ECE602, 3ECE624, 3ECM100, 3SFE617, 3ISE617, 3EEE612, 3SFE601, 3SFE611, 3MMC600, 3CCE633, 3ISE601, 3ECE621, 3CCE632, 3SFE615, 3MTS696, 3ECE616, 3MTS690, 3ISY651, 3ECE603, 3ISY612, 3IIS612, 3MTS694, 3ISE611	27	0	none	0	0	0.000000	0.000000	0.000000
41	3SFE602, 3ECE624, 3SFE620, 3MTS674, 3MMC600, 3SFE621, 3CCE632, 3SFE618, 3ECE602, 3ECM100, 3SFE617, 3ISE617, 3SFE610, 3EEE612, 3SFE605, 3MTS621, 3MTS691, 3SRT601, 3SFE601, 3SFE611, 3CCE633, 3ISE601, 3ECE621, 3MTS696, 3ECE616, 3SFE616, 3SFE609, 3ISE609, 3MTS690, 3ISY651, 3ECE603, 3ISY612, 3IIS612, 3MTS626, 3MTS694, 3ISE611	36	2	3MTS626, [3MTS691]	2	17,34	0.500000	0.027778	0.000000
42	3SFE622, 3ECE602, 3ISY657, 3SRT601, 3ISE601, 3SFE601, 3MMC600, 3ECE621, 3CCE632, 3MTS629, 3SFE615, 3ISY654, 3SFE609, 3ECE615, 3ISE609, 3ISY655, 3ISY651, 3SFE623, 3ISY608, 3SMT601, 3TSE613, 3TSE612, 3ISY612, 3IIS612, 3ISY652, 3SMC601, 3ISY699, 3IIS657, 3IIS655, 3IIS652, 3IIS651	31	0	none	0	0	0.000000	0.000000	0.000000
43	3MTS691, 3SFE623, 3SRT601, 3SFE630, 3MTS629, 3ECE624, 3MTS626	7	2	3MTS626 [3MTS691]	2	1,7	1.000000	0.285714	1.000000
44	3MTS691, 3SFE621, 3SFE602, 3ECE624, 3SFE610, 3SFE605, 3MTS621, 3SRT601, 3MMC600, 3CCE632, 3SFE616, 3SFE609, 3ISE609, 3MTS674, 3MTS626	15	2	3MTS626 [3MTS691]	2	1,15	1.000000	0.133333	1.000000
					Average:		**0.662267**	**0.371536**	**0.521739**
					MRR Without 'no answer' queries				0.705882

Study 2 SemaCS search (with personalisation): Year 2 Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 22 Prec				1.0				1.0			1.0
Q 23 Prec	0.0										
Q 24 Prec	0.0										
Q 25 Prec		1.0		1.0	1.0		1.0	1.0		1.0	0.0
Q 26 Prec				1.0				1.0			1.0
Q 27 Prec						0.1					0.1
Q 28 Prec						1.0					1.0
Q 29 Prec				1.0				0.7			0.8
Q 30 Prec											0.1
Q 31 Prec											1.0
Q 32 Prec			1.0	1.0		1.0		1.0	1.0		1.0
Q 33 Prec		1.0		0.7		0.8	0.8		0.0	0.0	0.0
Q 34 Prec						0.2					0.3
Q 35 Prec	0.0										
Q 36 Prec				1.0		1.0			1.0		0.0
Q 37 Prec											0.1
Q 38 Prec		1.0		1.0	1.0	1.0	1.0		0.0	0.0	0.0
Q 39 Prec	0.0										
Q 40 Prec	0.0										
Q 41 Prec						0.1					0.0
Q 42 Prec	0.0										
Q 43 Prec						1.0					0.3
Q 44 Prec						1.0					0.1

Study 2 SemaCS search (with personalisation): Year 2 interpolated average Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 22 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 23 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 24 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 25 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
Q 26 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 27 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Q 28 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 29 Prec	1.0	1.0	1.0	1.0	0.7	0.7	0.7	0.7	0.8	0.8	0.8
Q 30 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Q 31 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 32 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Q 33 Prec	1.0	1.0	0.8	0.8	0.8	0.8	0.8	0.0	0.0	0.0	0.0
Q 34 Prec	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Q 35 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 36 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
Q 37 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Q 38 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
Q 39 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 40 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 41 Prec	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0
Q 42 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 43 Prec	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Q 44 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.1	0.1	0.1	0.1	0.1
Average	0.52	0.52	0.51	0.51	0.50	0.50	0.46	0.38	0.38	0.34	0.29

Appendix B5 E: SRS search Year 2 P/R and F-score

Q-ID	Student Query description	Student Query	Expert generated query matches (ID)	SRS Generated	Matched #	Res Count	Match #	Recall	Precision	F-Score
22	applications	internet	3SFE617, [3ISE617, 3SFE615]	3ISE617, 3SFE617	2	2	3	0.667	1	0.8
23	programming language	ada	none	none	0	0	0	0	0	0
24	programming language	occam	none	none	0	0	0	0	0	0
25		business	3ISY612, [3IIS612, 3IIS655, 3ISY655, 3ECE615, 3ISY657, 3IIS657]	3IIS612, 3IIS655, 3ISY612, 3ISY655	4	4	7	0.571	1	0.727
26	I would like to know more about web design	Web Design	3SFE617, [3ISE617, 3SFE620]	none	0	0	3	0	0	0
27	i would like to know more about php and perl	internet programming	3SFE617, [3ISE617]	none	0	0	2	0	0	0
28		graphics	3MTS626, [3MTS691]	3MTS626	1	1	2	0.5	1	0.667
29	web designing	web designing	3SFE617, [3ISE617, 3SFE620]	none	0	0	3	0	0	0
30		website administration	3SFE615	none	0	0	1	0	0	0
31	i would like to learn more about flash	learn flash	none [3MTS691]	none	0	0	1	0	0	0
32		programming	3SFE605, 3SFE610, [3MTS621, 3SFE621, 3SFE602, 3ECE624]	3MTS621, 3SFE605,3SFE610	3	3	6	0.5	1	0.667
33	even though it is boring!	networks	3SFE611, [3ISE611, 3MMC601, 3SMC602, 3SFE622, 3TSE613, 3ISE619, 3SFE619]	3MMC601, 3SMC602	2	2	8	0.25	1	0.4
34	a platform games	java games	none [3MTS626, 3MTS690]	none	0	0	2	0	0	0

35	i would like to learn how to create a professional looking web site	html	none	none	0	0	0	0	0	0
36	javascript programme	java	3SFE605, [3SFE617, 3ISE617, 3MTS626]	none	0	0	4	0	0	0
37	help on actionscript	flash design	none [3MTS691]	none	0	0	1	0	0	0
38	i like to learn more about netwoks	network	3SFE611, [3ISE611, 3ISE619, 3MMC601, 3SFE619, 3SFE622, 3SMC602, 3TSE613]	3ISE611, 3ISE619, 3MMC601, 3SFE611, 3SMC602	5	5	8	0.625	1	0.769
39	graphic design	graphic design	none	none	0	0	0	0	0	0
40	I would like to learn more about database and its design	database design	none	none	0	0	0	0	0	0
41	learn more about 3d design	3d design	3MTS626, [3MTS691]	none	0	0	2	0	0	0
42	I'm interested in databases	database systems	none	none	0	0	0	0	0	0
43	3d, animation	computer graphics	3MTS626 [3MTS691]	none	0	0	2	0	0	0
44	designing games with multiplayer feature	graphics 3d multi	3MTS626 [3MTS691]	none	0	0	2	0	0	0

Appendix B5 F: SRS search year 2 interpolated average P/R and MRR

Study 2 UoW SRS search: Year 2 Recall Precision and MRR									
Q №	Query result	Result count	№ of Matches	Match IDs	Matches found №	Matched positions	Query Recall	Query Precision	RR/MRR
22	3ISE617, 3SFE617	2	3	3SFE617, [3ISE617, 3SFE615]	2	1,2	0.666667	1.000000	1.000000
23	none	0	0	none	0	0	0.000000	0.000000	0.000000
24	none	0	0	none	0	0	0.000000	0.000000	0.000000
25	3IIS612, 3IIS655, 3ISY612,	4	7	3ISY612, [3IIS612, 3IIS655, 3ISY655,	4	1,2,3,4	0.571429	1.000000	1.000000
26	none	0	3	3SFE617, [3ISE617, 3SFE620]	0	0	0.000000	0.000000	0.000000
27	none	0	2	3SFE617, [3ISE617]	0	0	0.000000	0.000000	0.000000
28	3MTS626	1	2	3MTS626, [3MTS691]	1	1	0.500000	1.000000	1.000000
29	none	0	3	3SFE617, [3ISE617, 3SFE620]	0	0	0.000000	0.000000	0.000000
30	none	0	1	3SFE615	0	0	0.000000	0.000000	0.000000
31	none	0	1	none [3MTS691]	0	0	0.000000	0.000000	0.000000
32	3MTS621, 3SFE605,3SFE610	3	6	3SFE605, 3SFE610, [3MTS621, 3SFE621,	3	1,2,3	0.500000	1.000000	1.000000
33	3MMC601, 3SMC602	2	8	3SFE611, [3ISE611, 3MMC601, 3SMC602,	2	1,2	0.250000	1.000000	1.000000
34	none	0	2	none [3MTS626, 3MTS690]	0	0	0.000000	0.000000	0.000000
35	none	0	0	none	0	0	0.000000	0.000000	0.000000
36	none	0	4	3SFE605, [3SFE617, 3ISE617, 3MTS626]	0	0	0.000000	0.000000	0.000000
37	none	0	1	none [3MTS691]	0	0	0.000000	0.000000	0.000000
38	3ISE611, 3ISE619, 3MMC601,	5	8	3SFE611, [3ISE611, 3ISE619, 3MMC601,	5	1,2,3,4,5	0.625000	1.000000	1.000000
39	none	0	0	none	0	0	0.000000	0.000000	0.000000
40	none	0	0	none	0	0	0.000000	0.000000	0.000000
41	none	0	2	3MTS626, [3MTS691]	0	0	0.000000	0.000000	0.000000
42	none	0	0	none	0	0	0.000000	0.000000	0.000000
43	none	0	2	3MTS626 [3MTS691]	0	0	0.000000	0.000000	0.000000
44	none	0	2	3MTS626 [3MTS691]	0	0	0.000000	0.000000	0.000000
					Average:		0.135352	0.260870	0.260870
					MRR Without 'no answer' queries				0.352941

Study 2 UoW SRS search: Year 2 Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 22 Prec				1.0				1.0			0.0
Q 23 Prec	0.0										
Q 24 Prec	0.0										
Q 25 Prec		1.0		1.0	1.0		1.0	0.0		0.0	0.0
Q 26 Prec				0.0				0.0			0.0
Q 27 Prec						0.0					0.0
Q 28 Prec						1.0					0.0
Q 29 Prec				0.0				0.0			0.0
Q 30 Prec											0.0
Q 31 Prec											0.0
Q 32 Prec			1.0	1.0		1.0		0.0	0.0		0.0
Q 33 Prec		1.0		1.0	0.0	0.0	0.0		0.0	0.0	0.0
Q 34 Prec						0.0					0.0
Q 35 Prec	0.0										
Q 36 Prec				0.0		0.0			0.0		0.0
Q 37 Prec											0.0
Q 38 Prec		1.0		1.0	1.0	1.0	1.0		0.0	0.0	0.0
Q 39 Prec	0.0										
Q 40 Prec	0.0										
Q 41 Prec						0.0					0.0
Q 42 Prec	0.0										
Q 43 Prec						0.0					0.0
Q 44 Prec						0.0					0.0

Study 2 UoW SRS search: Year 2 Interpolated Precision											
Recall-->	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Q 22 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0
Q 23 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 24 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 25 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
Q 26 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 27 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 28 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 29 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 30 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 31 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 32 Prec	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Q 33 Prec	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 34 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 35 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 36 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 37 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 38 Prec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
Q 39 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 40 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 41 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 42 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 43 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q 44 Prec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	0.26	0.26	0.26	0.26	0.22	0.22	0.13	0.04	0.00	0.00	0.00

Appendix B6: University of Westminster SRS study participant information

What is SemaCS?

SemaCS is an innovative search engine currently referencing School of Informatics module descriptions.

Why do you need my help?

Your help is needed to evaluate SemaCS, its accuracy and efficiency, as well as to identify any further possible improvements.

What is required of me?

You are to search for any School of Informatics free choice module (s) that you would like to take next year. Once you identify your module (s) – you will submit module ID (s) – and that's it!

How long would it take?

Only about 5 minutes of your time would be required!

Why would I want to donate 5 minutes of my time?

Many reasons really: you would be helping to drive science, you would get a chance to try SemaCS (and you would be one of the first people to do so, you get a nice warm feeling knowing that you have made a difference! And you get to help an unfortunate PhD student looking for test subjects ☺

What data about me are you going to store?

In simple terms – no personal data of any kind is to be collected and/or stored! We simply log your query, if you are first or second year undergraduate student, time taken and module ID (s) that you feel are a match to your query – that is all!

I am really interested – can I have the results? Ask questions?

Of course! Once study data has been analysed all results will be made available! And if you have any questions – do feel free to ask or to get in touch via sjachym@wmin.ac.uk (after you have finished searching though – we do not want you to be biased in any way!)