Search in the Eye of the Beholder: Using the Personal Social Dataset and Ontology-guided Input to Improve Web Search Efficiency

Juan Miguel Gomez[1], Ricardo Colomo Palacios[1], Giner Alor-Hernandez[2], Ruben Posada-Gomez[2], Angel Garcia Crespo[1]

[1]*Departamento de Informática*
Universidad Carlos III de Madrid, Spain
{juanmiguel.gomez, ricardo.colomo, angel.garcia}@uc3m.es

[2]*Division of Research and Postgraduate Studies*
Instituto Tecnologico de Orizaba, Mexico
{galor,rposada }@ itorizaba.edu.mx

Abstract

Among the challenges of searching the vast information source the Web has become, improving Web search efficiency by different strategies using semantics and the user generated data from Web 2.0 applications remains a promising and interesting approach. In this paper, we present the Personal Social Dataset and Ontology-guided Input strategies and couple them together, providing a proof-of-concept implementation.

1. Introduction

Vast information sources like the Web impose various challenges to browse or search using existing search engines, which accept minimal input in the form of keywords. Many strategies have been trying to increase the quality of information retrieval, from query expansion to the collaborative filtering or multifaceted browsing [1]. However, current approaches are still not fulfilling expectations, leading the user in many cases to frustration.

Recently, a new breed of user generated content aware technologies which have been encompassed by the "Web 2.0" buzzword umbrella have turned up to provide a huge amount of metadata and information about the user as a particular entity. Tags, picture sharing environments, social bookmarks, blogs and music preferences are just the tip of the iceberg.

In addition, semantic technologies are evolving to a more mature state in which ontologies [2], its backbone technology, provide a formal representation of a domain. The shift enabled by the use of machine understandable ontologies can outperform the current endeavors that require finding data spread out across the Web or dynamically drawing inferences which are continually hampered by their reliance on ad-hoc, task specific frameworks.

In this paper, we present two independent search and browsing strategies which enhance the efficiency in Web search, particularly in well-defined and concrete domains. Firstly, we describe the Personal Social Dataset (PSD), a strategy to gather the user generated metadata garnered by the aforementioned Web 2.0 technologies and turn it into a lightweight ontology which can be used for filtering search results. Secondly, we present an Ontology-guided Input tool, which provides query refinement and multifaceted browsing. These two approaches can be coupled together and we also present a proof-of-concept implementation on top a current Web search engine, additionally including the results of an evaluation in the search efficiency performed by the prototype.

The remainder of the paper is organized as follows. In section 2, we discuss the Personal Social Dataset as a basis for collaborative filtering [4]. In section 3, we describe our Ontology-guided Input tool as a means of semantically enhanced query refinement and its high efficiency in a particular dataset. In section 4, we describe how both approaches are coupled together and their breakthroughs for Web search efficiency. In section 5, we show our proof-of-concept implementation and the first preliminary results of the evaluation of our prototype. Section 6 spans over and

bind together a number of related works. Finally, section 7 concludes the paper and outlines our future work.

2. The Personal Social Dataset

Web 2.0 technologies as outlined in [3] are exemplified by *blogs*, namely easy to update websites about a particular subject where entries are written in chronological order, picture-sharing environments such as Flickr[1] or Photobucket[2], social bookmarking sites such as Del.icio.us[3], video-sharing such as YouTube[4] or music preferences such as Last FM[5]. A number of common features of Web 2.0 applications have been identified in [5]. First, community-awareness since Web 2.0 pages allow contributors to collaborate and share information easily. Secondly, services being pulled together into *mashups* in order to experience the data in a novel and enhanced way. Finally, AJAX as a technological pillar for building responsive user interfaces. The different sources of user generated data generated from Web 2.0 applications are depicted in the following figure.

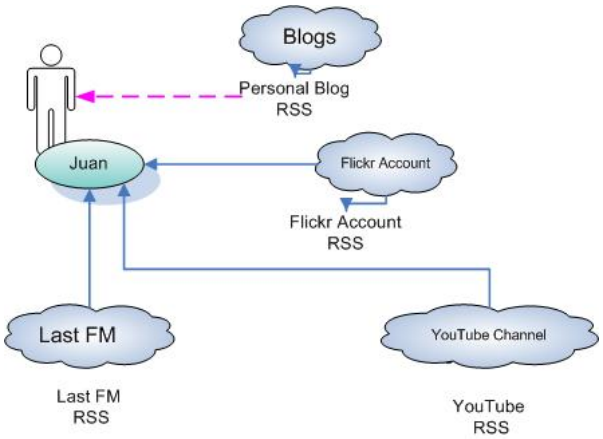

Figure 1. Data Sources from Web 2.0 Applications

However, another fundamental feature is the possibility of *tagging* the content in all these applications. Tags are freely chosen keywords describing a particular resource. They offer a simple way of retrieving content (e.g. retrieval of my pictures in Flickr with the tag *Norway*). These tag sets and their assignments to objects are envisaged as subjective conceptualizations, being potentially aggregated to a flat bottom-up categorization or folksonomy. In [6], Folksonomies have been claimed to be an interesting emergent attempt for information retrieval but serve different purposes to ontologies, the latter are attempts to more carefully define parts of the data world and to allow mappings and interactions between data held in different formats. Hence, ontologies are defined through a careful, explicit process that attempts to remove ambiguity, whereas the definition of a tag is a loose and implicit process where ambiguity might well remain. Finally, the inferential process applied to ontologies is logic based and uses operations such as join. The inferential process used on tags is statistical in nature and employs techniques such as clustering.

Nevertheless, in the past few years, there have been successful attempts of enriching tags with hierarchical relations [7] and the creation of faceted ontologies [8]. Furthermore, [9] describes the theory of formal classification, where labels are translated to a propositional concept language. Each node is associated to a normal formula that describes the content of the node, capturing g the knowledge that implicitly exists within simple classification hierarchies.

The Personal Social Dataset (PSD) is a lightweight ontology used for collaborative data filtering in which we follow an integrated approach of combining three types of techniques for improving its construction from the tag sets gathered from the aforementioned Web 2.0 user sources, namely: personal / professional blog, Flickr account, Del.icio.us account, YouTube channel and Last FM preferences. Most of these sources are syndicated in a RSS syntax, which we will use as main data streams.

The three techniques we are applying are as follows:

- Applying the Vector Space Model: The Vector Space Model [10] is an algebraic model used for information filtering, information retrieval, indexing and relevancy rankings. It represents natural language documents (or any objects, in general) in a formal manner through the use of vectors (of identifiers, such as, for example, index terms) in a multi-dimensional linear space. Documents are represented as vectors of index terms (keywords). The set of terms is a predefined collection of terms, for example the set of all unique words occurring in the

[1] Flickr: http//www flickr.com

[2] Photobucket: http://www.photobucket.com

[3] Del.icio.us: http://del.icio.us

[4] YouTube: http://www.youtube.com

[5] Last FM: http://www.lastfm.com

document corpus. Relevancy rankings of documents in a keyword search can be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as same kind of vector as the documents.

- Using Latent Semantic Analysis (LSA) [11] for analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA uses a term-document matrix which describes the occurrences of terms in documents. A typical example of the weighting of the elements of the matrix is the TF-IDF (Term Frequency–Inverse Document Frequency): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up-weighted to reflect their relative importance.
- Validating the set of terms pertaining to the PSD with online lexical resources, such as Wordnet[6]. Dictionaries are generally considered as a valuable and reliable source containing information about the relationships among terms (e.g. synonyms). Also Wordnet can add conceptual meaning to the tags and there is an RDF transcript available.

Fundamentally, the coupling of the three techniques strongly founded on the Information Retrieval literature roots provide a two-pronged approach to retrieve and accurate PSD: selecting and extracting the most accurate tags from the pool of Web 2.0 applications user generated content and creating "metadata cloud" which encapsulates the subjective meaning and intention the user conveyed through the tagging process. The PSD hence represents a valuable piece of knowledge which could be envisaged as a projection of the subjective mindset of the user.

3. Ontology-guided Input

The aim of Ontology-guided Input is to assist users in composing their sentences and search criteria. Most current search engines require a minimal set of keywords to retrieve a number of resources indexed, failing to organize and show the different relationships among those resources.

For this purpose, the Ontopath System[7] has been used as source of inspiration. Whereas Ontopath is implemented as a set of standalone Java applications, our system runs in the end-user web browser. Despite the intended functionality is the same as in Ontopath, our Ontology-guided Input (OGI) component differs in terms of scope and domain of search. Our OGI component offers the user a set of possible words to be inserted while writing in a particular statement. Specifically, it offers users different recommendations to complete the statements expressing their goals. These recommendations are based on knowledge extracted from an ontology repository. The ontologies of the repository contain the feasible actions. An example of an ontology-guided input can be spotted in the following figure.

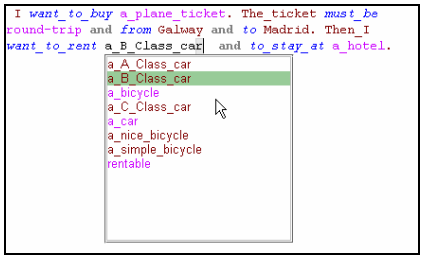

Figure 2. Example of an Ontology-guided Input

Our OGI comprises three tools:

- An Ontology Editor enables the creation of RDFS ontologies and individuals by means of a "classical" (lists and text-fields based) interface.
- A Graph Editor enables the visualization of individuals in a "nodes and arcs" style, and 3) Controlled Natural Language (DESC) enables the textual creation of individuals. We have imitated the functionality provided by DESC to allow end-users the creation of simple sentences to describe their goals. These sentences are constrained to valid sentences, i.e. sentences conform to a specific grammar depicted in Backus Naur Form (BNF) notation as follows:

```
Description::= Sentence { Sentence }
Sentence::= Phrase { "and" Phrase } "."
Phrase::= NounPhrase VerbPhrase{ "and"
VerbPhrase}
NounPhrase::= Subject | Subject "which"
VerbPhrase
VerbPhrase::= Predicate ObjectPhrase{
"and" ObjectPhrase}
```

[6] Wordnet: http://www.wordnet.com

[7] Ontopath: http://www.ontopath.com

```
ObjectPhrase::= Object | Object "which"
VerbPhrase
Subject::= Resource | "the" Resource |
"it"
Predicate::= Property
Object::= Resource | "the" Resource |
"it" | Literal
```

By examining this grammar it can be noted that it is built on three basic elements Resource, Property, Literal, which keep a tight relationship with their RDF counterparts, respectively Subject-Predicate-Object, also known in RDF syntax as triples [2]. Consequently, any sentence created in accordance with the depicted grammar has an immediate translation to RDF.

In this section, we have presented our OGI component and its grounding technical features. In the next section, we focus on integrating both the OGI component and the PSD approaches.

4. Filtering Web Search with PSD and OGI

The need to improve search efficiency by means of filtering strategies has been coming into increasingly sharp focus recently. Years ago, most data sat in silos attached to specific applications in legacy systems inside the boundaries of companies. Then the Web came into the arena, bringing the hurly-burly of data becoming available across applications, departments and entities in general.

However, throughout these developments, a particular underlying problem has remained unsolved: data reside in thousands of incompatible formats and cannot be systematically managed, integrated, unified or cleansed. To make matters worse, this incompatibility is not limited to the use of different data technologies or to the multiple different "flavors" of each technology (for example, the different relational databases in existence), but also because of its incompatibility in terms of semantics.

When using semantics as a searching technology, another problem that has been noted is the so-called "Semantic Web chicken-egg" problem [15]. The provider of the Web information or data source would always request for a good excuse or reason, a good application or benefit, from providing the metadata. However, if the metadata is not generated, no application or value-added functionality can be achieved.

In this work, we discuss the breakthroughs of using a combination of the PSD and the OGI coupled together as a means of improving Web Search results. The general flow of improvement coupling both strategies is shown in the next figure.

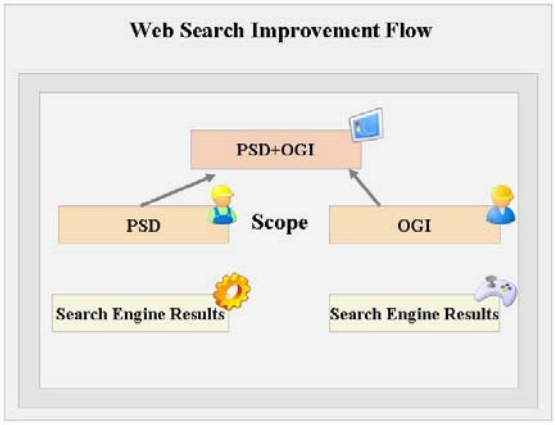

Figure 3. The PSD+OGI approach

Fundamentally, our architecture principles are based on the addition of both strategies. Firstly, the PSD is found out for the user and a traditional search engine results are filtered using the PSD correlation using again the Latent Semantic Analysis (LSA) [11] i.e. only those documents retrieved by the search engine with a particular high ranking correlation with the PSD are eligible to be transferred to the net phase. Secondly, the OGI filters a number of results and constitutes a basis for navigating through the graph of the PSD retrieving those results that have been selected by the previous phase in which the LSA metrics was held as a proximity measure.

This approach is better explained and discussed by means of an example, what is elaborated in the following section.

5. Implementation

We have built up our Personal Social Dataset Ontology guided Architecture (PODA), for short, prototype following the principles of the previous sections. Our implementation is built on top of typical current search engines, such as the Yahoo Search Engine or the Google Search engine.

Firstly, our prototype requests from the user a number of URIs directly related with its daily use Web 2.0 applications. Those URIs are searched and queried in order to find the related RSS feed. Once it has been crawled, the PSD Crawler retrieves the PSD belonging to the user. Secondly, the PSD is used in a two-pronged approach. If the domain search is reduced and it matches, the PSD is used as a basis for the Ontology Guided Input. Otherwise, it is simply used for the user to narrow down the search.

A screenshot of the PODA prototype is shown in the following figure.

Figure 4. The PODA prototype

Now let us depict a particular example of the implementation search and capabilities. Once the PSD has been retrieved about a particular user, John, who works in IBM, a number of concepts have been retrieved to configure the PSD which works as the backend of the Ontology-guided input. Let us suppose that this is an RDFS ontology with (1) classes Actor and Things. Class Actor has subclasses Enterprise and Human, and class Things has subclasses "rentable" (with subclass "a_car") and "buyable" (with subclass "an_appartment"), (2) properties want_to_buy (with domain Actor and range "buyable" and want_to_rent (with domain Actor and range "rentable", (3) individuals (instances in Object Oriented Programming jargon) "I" and "John" for class Human, "IBM" for class Enterprise. Given this ontology the Ontology-guided Input component allows guided creation of sentences such as:

```
{I/John/IBM} want_to_buy/rent}
{an_apartment/a car}
```

In this sentence, displayed options are shown between brackets, i.e. the first word can be selected from a list with the words "I", "John" and "IBM". The second word could be "want_to_buy" or "want_to_rent", and the third word should be either "an_apartment" or "a_car".

Some refinements could lead to the possibility to compose richer and more complex sentences. Some of these are:

- Temporal sequencing. Adding to the previous ontology the individual "Then_I" (from class Human) allow the creation of these phases:

```
I want_to_rent a car. Then_I
want_to_buy an_apartment.
```

- Giving details. Creating individuals of classes, for example "a_nice apartment" for class "an_apartment, and "a_C_class_car" for class "a_car", more detailed things can be expressed:

```
I want_to_rent a C_class_car.
Then_I want_to_buy
a_nice_apartment.
```

This process of specialization can be expanded with no limits to detail any Thing.

- Richer verbs. Subclassing properties allow richer verbs. For example, subclassing the property "want_to_buy" to create "wants_to_buy" expands the options list to enable the creation of this phrase:

```
IBM wants_to_buy  a_nice_apartment
```

- Domain specific sentences. In **¡Error! No se encuentra el origen de la referencia.**, a complex sentence related to "traveling" can be seen. To create this kind of domain-specific sentences, the ontology was populated with 1) classes "Ticket" (subclassing Actor and creating individual "The_ticket"), "Cities" (subclassing Things and with individuals "Galway" and "Madrid") and "AirplaneTicketDetails" (subclass of Things, and with subclasses "one-way" and "round-trip"), and 2) property "traveling" (with no domain and no range) and its subproperties "from" (domain Ticket and range Cities), "must_be" (domain Ticket and range AirPlaneTicketDetails). This is the example phrase:

```
I want_to_buy a_plane_ticket.
The_ticket must_be round-trip and from
Galway and to Madrid.
```

- Merging disparate domains. Using the previous ontology it is impossible to buy a rentable object, or to rent a buyable article. Let's say previous class "a_car". A possible solution is to create "a_car" as subclass of "rentable" (as it is now) and subclass of "buyable". This approach allows sentences such as:

```
I want_to_rent a_car and I want_to_buy
a_car.
```

Since the search results have been narrowed down dramatically, the efficiency of the approach is acknowledged.

6. Related Work

Since the work on improving search results spans over and binds together a number of research initiatives, in this section we briefly describe related work.

Searching has been subject of intensive research but a more concrete survey on filtering search results and optimizing results yields also a remarkable amount of efforts. Following research successfully implemented in the Google search engine [17], a number of search

variants related to the work presented have been explored such as using faceted search [20], including its application to multimedia faceted metadata for image search and browsing [18] or navigating RDF data [19].

Collaborative filtering was coined by Goldberg in [5] and it has been extensively used for data-intensive recommendation systems for personalized recommendations for music albums and artists as can be found in Ringo [12]. Active Collaborative Filtering solutions such as the one discussed in [13] focus on one-to-one recommendations and a social collaborative filtering system where users have direct impact in the final process is described in [14]. A similar work has been intended in SITIO, a Social Semantic Recommendation System [15], which combines the use of semantics with socially-oriented collaborative recommendation systems for the discovery and location of Web resources. Also, Semantic Social Collaborative Filtering has been used with FOAF [16].

A number of research initiatives by the authors of this work are related to combining Semantics with Web Services [22] or to software components [23].

7. Conclusions and Future Work

In this article, we have presented a novel approach to improve Web search results by adding PSD and Ontology-guided input. Particularly, we have discussed how these strategies can enhance search effectiveness in very concrete and well-defined domains.

In a larger context, the above-mentioned problem may be multiplied by thousands of data structures located in hundreds of incompatible databases and message formats. Actually, the problem is growing; the Internet exponential growth, also empowered by the amount of user generated data from Web 2.0 applications gathers data constantly, reengineer intense and massive data techniques processes and integrate with more sources. Moreover, developers are continuing to write new applications and to create new applications and databases based on requests from users, without worrying about overall data management issues.

Hence, our future work will consist of evaluating our implementation more carefully and look for case studies or datasets where pooling out of results can determine more accurately if the effectiveness of search takes place and how to follow such path.

8. References

[1] Kruk, S. Social Semantic Search and Browsing. DERI Technical Report. 2005.

[2] Fensel,D. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag. 2002.

[3] T. O'Reilly. What is Web 2.0 – design patterns and business models for the next generation of software. 2005.

[4] Ankolekar, A., Krötzsch, M., Tran, T., and Vrandecic, D. 2007. The two cultures: mashing up web 2.0 and the semantic web. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM Press, New York, NY, 825-834.

[5] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. Commun. ACM 35, 12. 1992.

[6] Shadbolt, N. Hall, W. Berners-Lee, T. The Semantic Web Revisited. IEEE Intelligent Systems. 2006.

[7] Schmitz, P. Inducing Ontology from Flickr Tags. Collaborative Web Tagging Workshop. Proceedings of the 15th WWW Conference. 2006.

[8] Heyman, P. Garcia-Molina, H. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report Stanford University. 2006.

[9] Giunchiglia, F. Marchese, M. Zaihrayeu, I. Towards a Theory of Formal Classification. Proceedings of the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications. Pittsburgh, Pennsylvania. 2005.

[10] Salton, G. Wong, A. and Yang, C. S. A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, nr. 11, pp 613–620. 1975.

[11] Deerwester, S. Dumais, Furnas, G. W. Landauer, T. K. Harshman, R. "Indexing by Latent Semantic Analysis". Journal of the Society for Information Science 41, Issue 6. pp 391-407. 1990.

[12] Shardanand, U. Maes, P. Social Information Filtering: Algorithms for Automating "Word of Mouth". In Proceedings of the ACM CHI95. 1995.

[13] Maltz, D. Ehrlich, K. Pointing the Way: Active Collaborative Filtering. In Proceedings of the Conference on Computer Human Interaction. 1995.

[14] Sugiyama, K. Hatano, K. Yoshikawa, H. Adaptive Web Search based on User Profile constructed without any effort. Proceedings of the WWW04. 2004.

[15] Gomez, J. M. Alor, G. Posada, R. Abud, A. Garcia, A. SITIO: A Social Semantic Recommendation Platform," Proceedings of the 17th International Conference on Electronics, Communications and Computers (CONIELECOMP'07). 2007.

[16] Kruk, S. Decker, S. Semantic Social Collaborative Filtering with FOAFRealm. In Proceedings of the Semantic Desktop Workshop. ISWC05. 2005.

[17] Bring, S. Page, L. The Anatomy of Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN systems. 30 (1-7), pp. 107-117. 1998.

[18] Yee, K.-P. Swearingen, K. Li, K. Hearst, M. Faceted Metadata for Image Search and Browsing. Proceedings of the International Conference of Human Interaction. 2003.

[19] Oren, E. Delbru, R. Decker, S. Extending Faceted Navigation for RDF Data. Proceedings of the International Semantic Web Conference (ISWC06). Athens, Georgia. 2006.
[20] Ranganathan, S. R. Elements of library classification. Bombay: Asia Publishing House. 1962.
[21] Schraefel, M. Wilson, A. Russell, and D. A. Smith. mSpace: Improving information access to multimedia domains with multimodal exploratory search. Communications of the ACM, 49(4), 2006.
[22] Gomez, J.M. Paniagua, F. García, A. Bussler, C. Modelling B2B Conversations with COOL for Semantic Web Services. Proceedings of the International Conference on Internet and Web Applicaitons and Services (ICIW06). Guadaloupe, France. Feb 19-25th, 2006.
[23] Gomez, J.M. Han, S. Toma, I. García, A. A Semantically-Enhanced Component-based Architecture for Software Composition. Proceedings of the International Multi-Conference on Computing in the Global Information Technology (ICCGI06). Bucarest, Romania. August 1-3rd, 2006.