Linking Textual Resources to Support Information Discovery

Thesis

Linking Textual Resources to Support Information Discovery

Petr Knoth

Submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

Knowledge Media Institute

The Open University

4^{th} November 2014

Milton Keynes

United Kindgom

Abstract

A vast amount of information is today stored in the form of textual documents, many of which are available online. These documents come from different sources and are of different types. They include newspaper articles, books, corporate reports, encyclopedia entries and research papers. At a semantic level, these documents contain knowledge, which was created by explicitly connecting information and expressing it in the form of a natural language. However, a significant amount of knowledge is not explicitly stated in a single document, yet can be derived or discovered by researching, i.e. accessing, comparing, contrasting and analysing, information from multiple documents. Carrying out this work using traditional search interfaces is tedious due to information overload and the difficulty of formulating queries that would help us to discover information we are not aware of.

In order to support this exploratory process, we need to be able to effectively navigate between related pieces of information across documents. While information can be connected using manually curated cross-document links, this approach not only does not scale, but cannot systematically assist us in the discovery of sometimes non-obvious (hidden) relationships. Consequently, there is a need for automatic approaches to *link discovery*.

This work studies how people link content, investigates the properties of different link types, presents new methods for automatic link discovery and designs a system in which link discovery is applied on a collection of millions of documents to improve access to public knowledge.

Acknowledgements

I am greatly thankful to my supervisor Zdenek. Zdenek always took the time to discuss my work when I asked for it. He was open to hear what I had to say before giving me his very constructive feedback. I truly enjoyed discussing my research with him. Zdenek never gave me tasks I had to do or prescribed articles I had to write, which I appreciated so much. This gave me the freedom to pursue the line of research I was interested in, while knowing I have the support of an experienced researcher in this area. Additionally, Zdenek has believed in my skills and the research I have been carrying out, which motivated me to continue. He gave me the courage to make hard decisions in situations when it was not clear what to focus on. For example, his support was essential in the decision to continue working on the CORE project after an initial six months long pilot, which substantiated the theoretical findings of this thesis, had been completed. At that time, it was not clear whether CORE will be able to make a difference, but Zdenek's trust in the need of this research and development effort was a crucial encouragement, which finally led to me having the privilege of seeing my work has picked up and has been built upon.

I also highly appreciated the help of my second supervisor Trevor. I always admired Trevor's "positive obsession"/perfectionism in organising research lit-

erature. He was the first researcher I have met able to find a printed and annotated version of an article he read years ago in minutes. While I was inspired by this, I have to admit I did not manage to meet this level of performance even with the use of computer based article reference managers. Trevor has an extensive knowledge of various requirements and conventions in structuring and writing research articles. This proved to be very useful in my PhD as Trevor helped me by reviewing research papers I submitted to conferences and by commenting on the thesis in the final stages of the writing process.

I would like to thank my colleagues with whom I had the opportunity to work on topics related to the research presented in this thesis. They are primarily Jakub Novotny, Lukas Zilka and Drahomira Herrmannova. Their enthusiasm in my work and this strand of research has been a great motivation. Additionally, I should not forget the current and previous members of my team, which I was lucky to establish at later stages of my PhD thanks to the funding for CORE I received from Jisc, AHRC/ESRC, NWO, Open University and the European Commission. They include Lucas Anastasiou, Vojtech Robotka, Samuel Pearce, Magdalena Krygielova, Matteo Cancellieri, Tomas Korec and Gabriela Pavel. CORE would not be as far as it is now without their help. In this respect, I would also like to express my gratitude to a few people in Jisc who were responsible for managing the funding CORE received. They recognised the potential and impact of this work and provided subsequent

funding for it. They include Balviar Notay, Neil Jacobs, Stuart Dempster and Andy McGregor.

I am also thankful for the opportunity to carry out this work in KMi. It has been a privilege to be able to work in such a great lab, be able to attend conferences and meet people who made me passionate about research. I also greatly appreciated a seemingly minor thing of being moved to an office from the open space after spending four years in KMi. While it was possible for me to do routine administrative work in the open space, it is still hard for me to understand how are people capable of producing great research papers in this way. The move to an office had a huge impact on the quality of work I could do and I am very sorry for the people who were not as lucky as me.

Last but not least, I would like to acknowledge my family. My mother and father always encouraged me to study and were supportive of me doing a PhD abroad. My wife Mirka left her family and friends in the Czech Republic and joined me after finishing her studies to start a new life in the UK. She has been a great support to me throughout my whole PhD, but especially in the final year by helping me to find the time and the mental strength to rewrite all the work I have done again as a single document. A huge motivation to finish this work provided also my son Tobias. I so much love spending time with him that I did not want to miss his childhood by not finishing the writing soon :) I am really indebted to him for this stimulus.

Publications

The research work presented in this thesis was used to produce the following research papers:

- Knoth, P. and Herrmannova, D. (2013) Simple Yet Effective Methods for Cross-Lingual Link Discovery (CLLD) - KMI @ NTCIR-10 CrossLink-2, NTCIR-10 Evaluation of Information Access Technologies, Tokyo, Japan

- Knoth, P. and Zdrahal, Z. (2012) CORE: Three Access Levels to Underpin Open Access, D-Lib Magazine, 18, 11/12, Corporation for National Research Initiatives

- Knoth, P. (2013) CORE: Aggregation Use Cases for Open Access, Demo at Joint Conference on Digital Libraries (JCDL 2013), Indianapolis, Indiana, United States

- Knoth, P., Zilka, L. and Zdrahal, Z. (2011) KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis, NTCIR-9: Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Tokyo, Japan, pp. 495-502

- Knoth, P., Zilka, L. and Zdrahal, Z. (2011) Using Explicit Semantic Analysis for Cross-Lingual Link Discovery, Workshop: 5th International

Workshop on Cross Lingual Information Access: : Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011), Chiang Mai, Thailand Publications — Download PDF Publications — Visit External Site for Details

- Knoth, P. and Zdrahal, Z. (2011) Mining Cross-document Relationships from Text, The First International Conference on Advances in Information Mining and Management (IMMM 2011), Barcelona, Spain

- Knoth, P., Novotny, J. and Zdrahal, Z. (2010) Automatic generation of inter-passage links based on semantic similarity, The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China

The research ideas presented in the thesis developed from or contributed to the following publications:

- Knoth, P., Herrmannova, D. (2014) Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing Research Contribution, In 3rd International Workshop on Mining Scientific Publications (WOSP 2014), Digital Libraries (DL 2014), London, United Kingdom

- Knoth, P., Anastasiou, L. and Pearce, S. (2014) My repository is being aggregated: a blessing or a curse? Open Repositories 2014, Helsinki,

Finland

- Knoth, P. (2013) From Open Access Metadata to Open Access Content: Two Principles for Increased Visibility of Open Access Content, Open Repositories 2013, Charlottetown, Prince Edward Island, Canada

- Knoth, P., Zdrahal, Z. and Juffinger, A. (2012) Special Issue on Mining Scientific Publications, D-Lib Magazine, 18, 7/8, Corporation for National Research Initiatives

- Herrmannova, D. and Knoth, P. (2012) Visual Search for Supporting Content Exploration in Large Document Collections, D-Lib Magazine, 18, 7/8, Corporation for National Research Initiatives

- Knoth, P., Robotka, V. and Zdrahal, Z. (2011) Connecting Repositories in the Open Access Domain using Text Mining and Semantic Data, International Conference on Theory and Practice of Digital Libraries 2011 (TPDL 2011), Berlin, Germany

- Knoth, P., Collins, T., Sklavounou, E. and Zdrahal, Z. (2010) Facilitating cross-language retrieval and machine translation by multilingual domain ontologies, Workshop: Workshop on Supporting eLearning with Language Resources and Semantic Data at LREC 2010, Valletta, Malta Publications

- Knoth, P., Collins, T., Sklavounou, E. and Zdrahal, Z. (2010) EUROGENE: Multilingual Retrieval and Machine Translation applied to Human Genetics, Demo at 32nd European Conference on Information Retrieval (ECIR 2010), Milton Keynes, United Kingdom

- Fernandez, M., Sabou, M., Knoth, P. and Motta, E. (2010) Predicting the quality of semantic relations by applying Machine Learning classifiers (Best Poster Award), Poster at EKAW 2010 - Knowledge Engineering and Knowledge Management by the Masses, Lisbon, Portugal

- Knoth, P. (2009) Semantic Annotation of Multilingual Learning Objects Based on a Domain Ontology, Workshop: Doctoral consortium at Fourth European Conference on Technology Enhanced Learning (ECTEL 2009), Nice, France

- Zdrahal, Z., Knoth, P., Collins, T. and Mulholland, P. (2009) Reasoning across Multilingual Learning Resources in Human Genetics, ICL 2009, Villach, Austria

- Knoth, P., Schmidt, M., Smrz, P. and Zdrahal, Z. (2009) Towards a Framework for Comparing Automatic Term Recognition Methods, Znalosti 2009, Brno, Czech Republic

Invited Talks

The following invited talks, discussing the research presented in this thesis, were given during the course of this work.

- Knoth, P. (2014) Discovering the Power of Open Access Scholarly Data – The Challenges and Opportunities. International Workshop on Challenges & Issues on Scholarly Big Data Discovery and Collaboration associated with IEEE BigData 2014, Washington D.C., United States.

- Knoth, P. (2014) Linking Open Access Resources to Improve Access to Public Knowledge. Open Access Week, Open University Library, Milton Keynes, United Kingdom.

- Notay, B., Jacobs, N., Muriel, M., Hubbard, B., Knoth, P., McIntyre, R. (2014) Jisc: Building a Cohesive Repository Shared Services Infrastructure for the UK, Open Repositories. Helsinki, Finland. Invitation to panel.

- Knoth, P. (2014) CORE: Aggregating and Enriching Repository Content to Support Open Access. Czech Open Repositories (Otevřené Repozitáře 2014). Brno, Czech Republic.

- Knoth P., (2013) CORE – Opening Up Content from Institutional Repositories. Annual meeting of the European Library 2013. Amsterdam, The

Netherlands.

- Knoth, P. (2013) From Open Access Metadata to Open Access Content: Towards an Infrastructure for Mining Scientific Publications. 2nd International Workshop on Mining Scientific Publications (WOSP 2013) associated with Joint Conference on Digital Libraries (JCDL 2014). Indianapolis, United States.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The amount of new information accessible online is increasing rapidly. With the growing amount of new data and the speed with which it is being updated comes the need to organise the data both effectively and efficiently, to ensure the information contained in the data can be accessed and reused when needed.

By "effectiveness of organising information", we understand the property of the data representation, which makes it possible to efficiently exploit the expressed information in the context of a given use case. For example, we can talk about an effective way of organising text documents for the purposes of keyword search. The *inverted index* structure (Manning et al., 2008) might be in this context considered an effective data representation. Similarly, one can talk about an effective data representation for the purposes of searching books or online shopping.

By "efficiency of organising information", we understand the property of a method, which enables us to transform input data into an effective representa-

tion, in time and with resources reasonable in the context of a given use case. A practical way to measure efficiency might be (in some context) time and space complexity. One can, for instance, compare the efficiency of different data indexing strategies for the purposes of finding related information. In the same way, it is possible to discuss efficient strategies for sorting books for the purposes of looking up their titles or strategies for entering shopping items into a database used for online shopping.

Designing effective yet efficient solutions for organising information at a Web scale is becoming increasingly challenging due to the growing amount of information, the need of exploiting it and the variability of situations in which we use it. Consequently, the current information seeking solutions, including those provided by the major search engines, are challenged by a substantial demand for information seeking experiences that are different from traditional content look up. They are according to Marchionini (2006) referred to by the term *exploratory search* and encompass activities, such as discovering, analysing, comparing or interpreting.

In this thesis, we will focus our attention on a type of an investigative exploratory search that can be regarded as *discovery* and is concerned with detecting links or relationships between resources. The research area concerned with developing solutions to this problem does not have a single established name. In the literature, it can be referred to as *link discovery*, *link generation*,

link detection, *relation extraction* or *content recommendation*. We will see that there are often major differences in the way this task is understood, defined, interpreted, applied and evaluated by different research groups. As a result, systems, referred to as link discovery systems, might be effectively solving very diverse problems.

In this work, we will concentrate on textual data. Link discovery methods and tools in this area build primarily on information retrieval, natural language processing and discourse analysis. By link discovery, we understand the task of linking documents or textual fragments of a lower granularity with respect to a given criterion of their semantic relatedness.[1] These links can be used, in turn, to improve the effectiveness of discovery and navigation in large textual collections. Link discovery systems typically need to process very large volumes of data, meaning the process of discovering links must be efficient and therefore as much automated as possible.

In Chapter 2, we will discuss link discovery in a wider context and see how it relates and compares to other problems of effective and efficient data organisation. We will then analyse the gap and formulate the research questions. In Chapter 3, we will formally define the link discovery task, review the state-of-the-art in link discovery and discuss approaches to the evaluation of link discovery methods. In Chapter 4, we will create a simple link discovery system based on measuring semantic similarity and explore its relation to the

[1] See Section 3.1 for a formal definition.

way people link content. Chapter 5 follows by investigating the relationship between semantic similarity and the qualitative properties of links. Chapter 6 introduces new methods for cross-lingual link discovery and explores how closely these methods simulate the performance of humans on the same task. Chapter 7 presents the link discovery methods we designed and submitted to two evaluation conferences, their results and comparison with other teams. It also explores issues in the evaluation methodology used at the evaluation conference and suggests improvements. Chapter 8 demonstrates our effort to improve access to public knowledge by creating a large-scale aggregation system for research papers with integrated link discovery methods. Finally, Chapter 9 summarises the contribution of the thesis and concludes.

Chapter 2

Motivation

2.1 Motivating scenario

Let us now introduce the research problem with a motivating example. Consider the visionary article *The Semantic Web* by Berners-Lee et al. (2001) at the time it was written. The article argued for a different World Wide Web than the one that existed at that time. It tried to convince the reader of the opportunities of moving beyond the traditional Web, which had been designed for people to read, to a new Web, where machines could understand and interpret information. When the article was written, it was submitted to Scientific American, where it was published on May 17th, 2001.

Let us assume that, before the article was published, the editor had asked the authors for metadata. They included keywords, the relevant research areas according to a given taxonomy and references to related information sources. The article together with the metadata was then made available on the Inter-

net. When the article was published, the role of the metadata was to make it easier to retrieve the article, discover it, and explore related resources. While we do not know which metadata were used at that time, it is reasonable to assume that the authors could use the free-text keyword *semantic web* and that they could associate the article to classes of a formal taxonomy, such as *knowledge representation*, *multi-agent systems* and *ontologies*. The original article does not mention or cite any related articles, but it is fair to say that a specific set of related articles could have been selected at the time of publishing.

Since the article was published, many new papers discussing this topic have appeared and numerous research studies have been carried out. This information growth had an impact on the quality/suitability of the metadata that were specified. In particular, the domain and its terminology have evolved. For example, the popularity of the term semantic web has steadily decreased over the last few years (see Figure 2.1). An extremely related term *internet of things*, which could have also been used as a keyword in 2001 had it been widely known, has become very popular instead (see Figure 2.1).

The number of articles discussing semantic web has dramatically increased since 2001 (and especially in the first few years following 2001) when this paper was published, this has been reflected in formal taxonomies, which now contain the class *semantic web* (in 2001 this was only a free-text keyword as the field had not been established yet). We could probably say that this

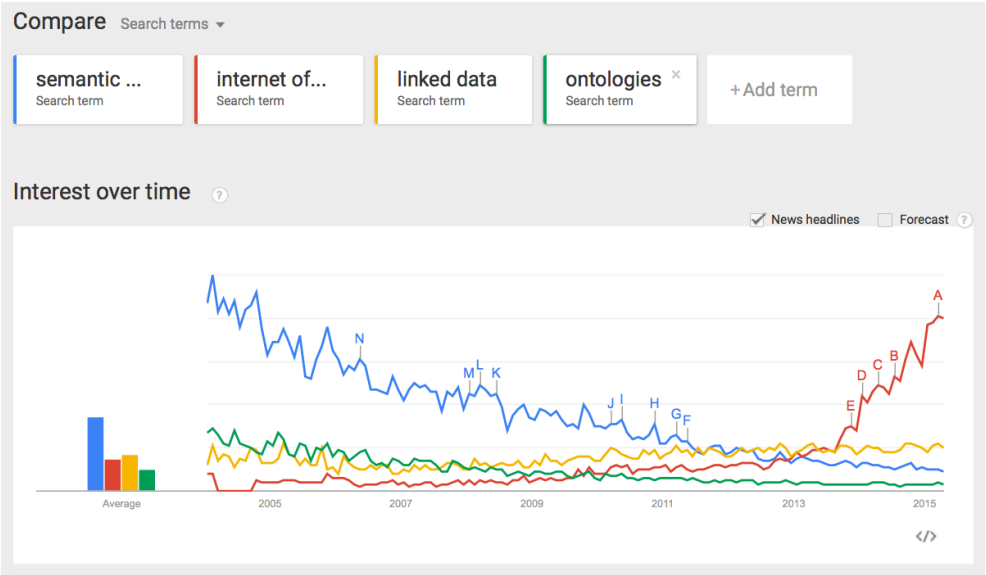


Figure 2.1: The popularity of topics in time (the figure is acquired from Google Trends) indicating that the best possible metadata description at a specific time might not be the best after a certain period.

article can also be seen as one that provided motivation for a new formal class, *linked data*, another closely related topic, which has recently and interestingly become slightly more popular than semantic web itself.

Finally, we can also see that new articles which are closely related to this original paper have been written and published. These might include *Linked Data: The Story So Far* by Bizer et al. (2009) as well as the *Internet of Things* by Kopetz (2011). To aid exploration, it would be useful for readers to be able to navigate from the semantic web paper to these. While existing search engines allow one to navigate to the former article thanks to a reverse citation link, navigating to the latter is not obvious as no citation has been explicitly stated.

2.2 Research problem

The motivating scenario indicates that keeping information organised, so that it remains well accessible over time is a challenging task. Accessibility is an abstract concept originally used in urbanism (Hansen, 1959) defined as a measure of potential opportunities for interaction with resources, such as employment, schooling or dining. Azzopardi and Vinay (2008) use an analogy of this definition to define *accessibility* in the context of information retrieval as *the potential of documents for retrieval*.

Based on this definition, we can say that in a hyper-linked collection exposed by a browsing system, a page with no incoming links will have no accessibility, while a page with thousands of incoming links will be very accessible. Another way to look at accessibility is to consider it in the context of a retrieval system. Azzopardi and Vinay (2008) proposed to measure the accessibility of document d given a retrieval system according to the following definition:

$$A(d) = \sum_{q \in Q} o_q . f(c_{dq}, \theta) \tag{2.1}$$

where o_q denotes the likelihood of expressing a query q from the universe of queries Q and the $f(c_{dq}, \theta)$ is a generalised utility/cost function where c_{dq} is the distance associated with accessing d through q which is defined by the rank of the document, and θ is a parameter or a set of parameters given the specific

type of measure. For example, if $\theta = c$, where c denotes the maximum rank that a user is willing to proceed down the ranked list, the function $f(c_{dq}, \theta)$ might return the value of 1 if $c_{dq} \leq c$ and 0 otherwise. Alternatively, the function $f(c_{dq}, \theta)$ can reflect the effort associated with finding the link to a document, etc.

Using this formalism, we can see that resources with high quality metadata will typically have higher *accessibility* than resources with poor metadata, because they are more likely to be retrieved and receive a lower rank in response to a query q. Therefore, high quality metadata are decreasing the distance associated with accessing d.

Since the time a resource is made available, the accessibility of that resource typically deteriorates over time (Figure 2.2). This accessibility deterioration can be prevented by applying forces aiming at improving or updating the resource's metadata. The magnitude of the accessibility deterioration is largely dependent on two variables: (1) the type of the information seeking behaviour (or more precisely the type of metadata needed to support the information seeking behaviour) and (2) the magnitude of the accessibility deterioration forces, which is influenced by a number of aspects including the frequency of content additions or updates in the collection as well as the speed of progress in the field described by the collection.

If we were developing a model of the accessibility deterioration forces, we

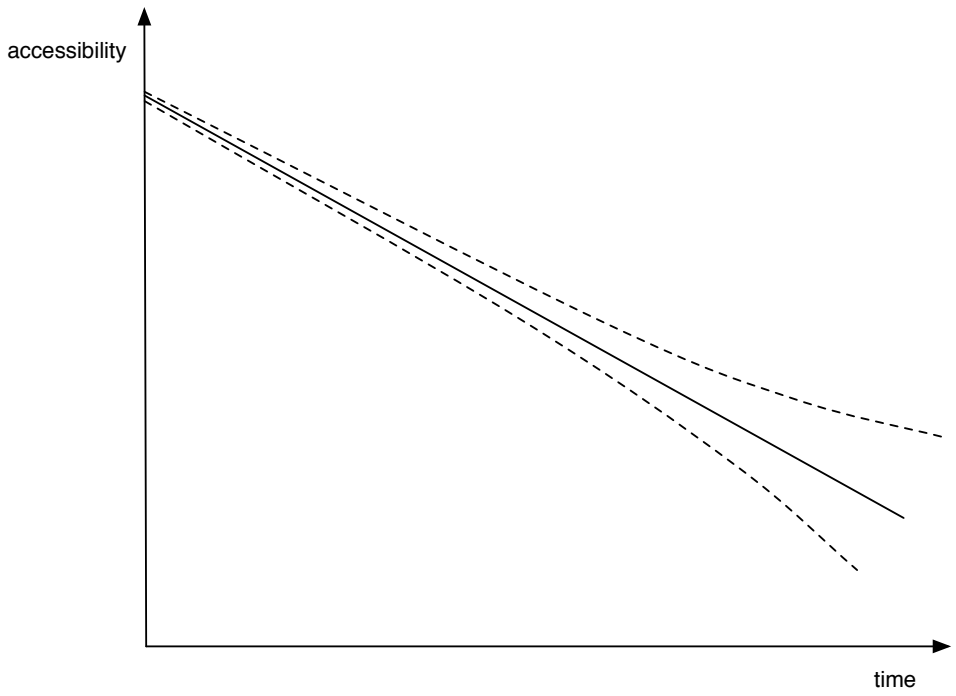


Figure 2.2: Accessibility deterioration over time. The exact type of the relationship (linear or hyperbolic) depends on external circumstances, such as the dynamics of change in the use of terminology and the changes in the popularity of the domain itself.

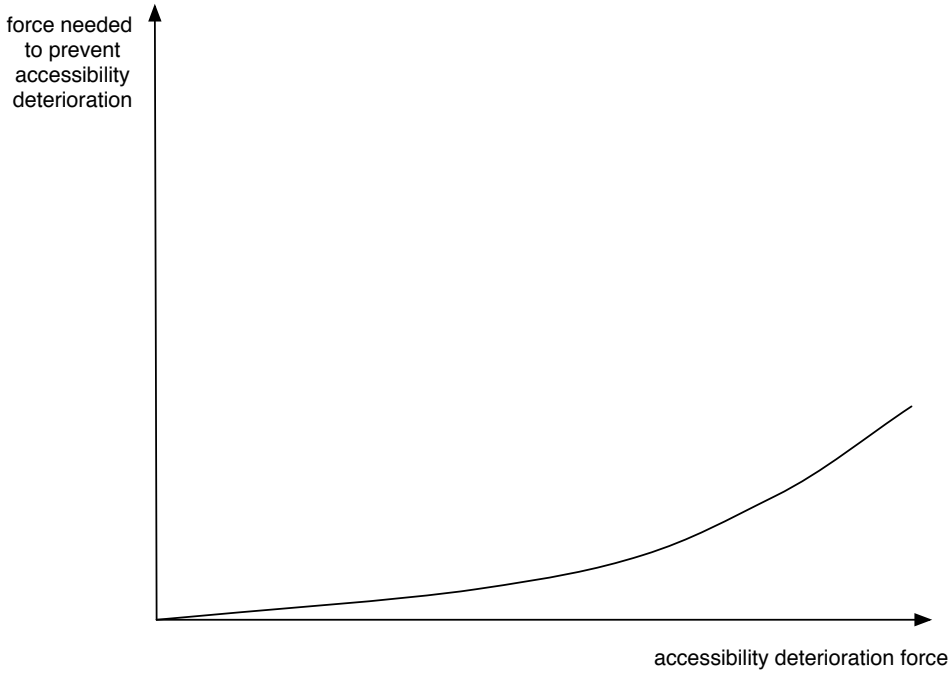


Figure 2.3: The force needed to prevent accessibility deterioration rises faster than linearly with respect to the accessibility deterioration force.

could expect that the two mentioned variables would likely be its main components. It is possible to see that such model is not necessarily linear (Figure 2.3), meaning that the amount of accessibility deterioration forces is not necessarily directly proportional to the amount of the forces needed to prevent the deterioration. For example, this type of non-linear behaviour was observed in the context of the Wikipedia project (Kittur et al., 2007). It is reported that the effort (force) necessary for the maintenance of the information in Wikipedia is not directly proportional to the amount of information stored, but rises faster than directly proportional to the amount of information.

So, how can we do better than in the motivating scenario? First of all, we

need to be aware of the information seeking behaviour we want to support, understand the types of metadata that are needed to support this behaviour and be realistic about the effect of accessibility deterioration forces on the metadata. Once this is done, we should focus our attention on designing effective yet efficient methods that can support the organisation of resources for this type of information seeking behaviour. As we are focusing our attention on exploratory information seeking (in particular, the discovery of links between resources), let us now present how we would envisage this to work on the example provided at the beginning of this section.

Since the article by Berners-Lee et al. (2001) was published, many new papers discussing similar issues have appeared and many new research studies have been carried out. As new terminology evolved, such as *linked open data* or *internet of things*, topic taxonomies have changed, for example the topic *semantic web* has been established, and new relevant sources of information appeared, the metadata of our article were automatically updated to reflect these changes. This has made the resource richer and increased its accessibility. The resource has not only been constantly easily accessible through general keyword search systems using contemporary terminology, but even people who have been browsing through categories of their interest have been able to easily find it. Furthermore, any time new semantically relevant content was made available on the Web, it was linked with our resource. These links do not nec-

essarily commence and point to the whole resource, but sometimes link only from/to a concept or a paragraph in the relevant resource. For example, there is now a link from *Internet of Things* by Kopetz (2011) to the first paragraph of *The Semantic Web* by Berners-Lee et al. (2001) where it is described that upon receiving a call, a phone sends a message to turn down the sound to all local devices that have a volume control. These links allow discovering relevant and sometimes, from the perspective of the reader, unexpected content, which one would not be able to easily find using keyword search. Why? Because one would not be able or would not imagine to formulate such queries. Consequently, these links are improving the accessibility of our resource, as defined in Equation 2.1, by increasing the likelihood o_q of expressing certain queries. The system we are interacting with also gives us the opportunity to see how other content is relevant to our resource. For example, we can see that (a) three other authors also published papers about the same topic, but analysed the problem from a broader perspective, (b) a number of new subsequent political studies, government policies and articles build on the ideas about the reduction of CO_2 emissions originally mentioned in our article, (c) a few new in-depth studies talking about the pitfalls of the prediction model we used appeared, (d) the results of a certain group of researchers who are using a different modelling approach seem to contradict our findings, etc.

There are many analogous situations and application domains in which one

benefits from discovery experiences. For example:

- *News reading* — A person reading newspapers wants to be informed about related news articles on the topic of interest. These include preliminary and subsequent news, new articles on the same topic written from different angles by different journalists and published in a variety of newspapers, various news analyses on the topic from a selection of critics/analysts, etc.

- *Employee training* — a new employee in a company can face the task of getting familiar with the business processes and workflows documented on the company's Intranet. She needs to discover and understand those that are relevant, but does not have time to read all available information.

- *Education in life sciences* - A student of medicine doing a introductory course in genetics needs to understand the main genetic concepts and their relationships. At the same time, an experienced medical professional wants to keep track of newly discovered relationships between genes and diseases as this might influence the advice he gives to his patients.

- *Arts and history* — An arts museum visitor who became excited about a certain topic, such as impressionism, during their last museum visit,

wants to find out interesting information about the life of modern impressionists.

All of the mentioned examples have something in common. The nature of the information seeking task makes it difficult to formulate suitable queries. In many cases, it would also require repetitive submission of these queries as it targets newly provided, so-called "fresh," information. It requires a passive rather than an active approach, meaning that "I prefer to be told" rather than "I want to ask."

A nice way how to demonstrate the importance of discovery techniques is by borrowing the terminology from the following quote:

> "There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say, we know there are some things we do not know. But there are also unknown unknowns the ones we dont know we dont know." — United States Secretary of Defense, Donald Rumsfeld.

Perhaps the main reason why discovery techniques are important is that there is one more group not mentioned in the quote — the *unknown knowns* — that is the things we don't know that are known. This is where discovery techniques are needed.

As opposed to look-up information seeking, it can be seen that discovery requires the knowledge of the relationships between resources rather than just

the knowledge of the resources. Consequently, to support discovery, we need metadata describing the relationships between resources, rather than just the resources.

Humans make use of relationship information all the time. Since we are born, our brain starts learning associations (relationships). Many of these associations (especially in our childhood) are learned by observing (Bandura et al., 1961) and interacting with the environment (Bruner, 1961) — the information is discovered in the information retrieval sense. In this way, we often learn as a result of coming across information which triggers our learning curiosity and desire rather than by pro-actively seeking that information in the first place. While the mainstream way of information seeking on the Internet is by submitting keyword queries to search engines, this approach limits the type of information we can be confronted with to content on topics about the existence of which we are aware. So, how can we better support discovery of information on the Web?

In fact, the way people access information on the Web has been over the last few years rapidly evolving towards incorporating more discovery experiences. For example, the ability to *follow* people or topics on social networks and receive new information about them immediately as it appears online is an example of such experience. The recommendation tools, which are currently being adopted by online retailers, Internet radios, media organisations and

also libraries, are becoming almost ubiquitous. These discovery tools are not a substitute for traditional search, they are rather essential tools complementing them.

In this chapter, we will explore how relationship metadata can be acquired and represented in large collections of textual documents. We will focus our attention on those relationships that are not present within the boundaries of a document, but rather exist across documents. We will call these relationships *cross-document relationships*. We will start by analysing the theoretical characteristics of different metadata types and discussing their implications for content organisation (Section 2.3,). Section 2.4 reviews existing approaches for acquiring different types of metadata and discusses approaches which depend on human curated organisation of content as well as those supported by technology. Focusing on relationship information, which is vital in information discovery, we will outline (in Section 2.5) various types of textual relationships and discuss more closely the types of relationships important in the context of the research task. In Section 2.6, we will provide a short discussion of the benefits and drawbacks of metadata standardisation efforts that are essential for organising information on the Web. Finally, we will specify the objectives of the thesis in Section 2.7.

2.3 Acquiring metadata

Information retrieval systems typically do not only index the data of resources, but make also use of the resources' metadata. These metadata generally improve the accessibility of the resource by providing additional information to the retrieval system. There are different types of metadata, with different properties. We identify three main classes of metadata types:

Type 1 — *Metadata describing the content of a resource.* This type is used to specify information to identify the resource or its part, such as the author's name, title of a resource or a set of free text keywords relevant to the resource.

Type 2 — *Metadata classifying the resource.* This type is used to associate the resource or its fraction with an external structure. For example, to categorise the resource according to a taxonomy of the subject domain or to semantically annotate/tag certain parts of the resource with respect to an external knowledge structure/ontology. For example, this metadata can express the classification of the resource (or the tagging of the resource with keywords) using the Dewey Decimal[1] or ACM classification[2].

Type 3 — *Relationship metadata.* This type of metadata can be used to de-

[1] `https://www.oclc.org/dewey.en.html`
[2] `http://www.acm.org/about/class/`

fine a (semantic) relationship between resources, such as links to related resources (e.g. `dc:relation`) or supporting materials (e.g. citations).

All of the three metadata types can be found in a number of widely adopted standards that aim at providing interoperability between systems. We will discuss them in Section 2.6. We will now analyse each metadata type with respect to the information needed for its creation, the time complexity of creation and durability.

2.3.1 Information needed for metadata creation

Type 1 metadata can often be extracted from the knowledge contained in a given resource.[3] The creation of Type 2 metadata requires, in addition, an understanding of some external structure. Providing Type 3 metadata requires one to consider (in the most extreme case) the relationship of the given resource to all other resources.

2.3.2 Time complexity of metadata creation

We will now explore the maximum time complexity for the provision of the previously mentioned metadata types. We assume here the metadata are created by a human, but the expressions below analogically apply also to software

[3]Usually, the creation of Type 1 metadata involves, in addition, the registration of the resource and its association with a unique and persistent identifier.

generated metadata. We will provide more details about the differences at the end of this section.

Let t_1 denote the maximum time needed to access, read/broadly understand a resource. Let h denote the number of concepts in a taxonomy and let t_2 denote the maximum time needed to check whether a given resource should be associated with a node in the taxonomy. Finally, let n be the number of resources available in a collection. Then, the maximum time t_{max} needed to provide Type 1 metadata for all items in the collection is:

$$t_{max} = t_1 n \Rightarrow t(n) = O(n), \tag{2.2}$$

thus, the time complexity is linear with respect to the number of resources in the collection. For simplicity, we assume appropriate metadata annotation can be provided in no time once the resources has been read/understood.

The maximum time needed to generate Type 2 metadata is:

$$t_{max} = (t_1 + t_2 h)n \ \Rightarrow t(n) = O(n). \tag{2.3}$$

The maximum time is given by the time of accessing, reading/understanding a resource plus the time of annotating the resource with respect to a taxonomy, this all times the number of resources available in the collection. The complexity is still linear with respect to the number of resources available, but

the actual time required for annotation rises also linearly with respect to the number of concepts in the taxonomy. We assume a resource can be associated to any number of the taxonomy concepts. In cases where just one taxonomy concept from a tree shaped structure is selected, the maximum cost of association is logarithmic with respect to the number of taxonomy concepts. The base of the logarithm is given by the branching factor of the tree structure.

Finally, the maximum time spent in deriving Type 3 metadata is given by the following expression:

$$t_{max} = (t_1.n).[t_1.(n-1)] \Rightarrow t(n) = O(n^2) \tag{2.4}$$

The equation states that for the generation of links specifying one type of a binary semantic relation it is necessary to access all resources and to take into account all remaining resources for each of them. The time complexity is thus quadratic with respect to the number of all available resources. For simplicity, the equation assumes that the annotator does not have any memory and thus needs to access, view and understand a particular resource each time again. While it is possible, in practice, that an annotator (with a memory) can generate metadata much faster than in the maximum time predicted by the equation, it is not possible to avoid the quadratic number of comparisons with respect to the number of resources. In addition, in very large collections, the ability of an annotator to keep a substantial part of the collection in memory

can be limited.

2.3.3 Metadata durability and the 3C rule

In real-world collections, new resources are often being added to the collection over time as demonstrated in the example at the beginning of this chapter. As new resources are being added, it is important that the collection metadata keep satisfying the following three conditions:

- *Consistency* — All metadata describing resources within the collection should be created conceptually using the same approach.
- *Completeness* — All metadata fields for describing resources in the collection should be populated (if that metadata field is applicable).
- *Correctness* — The addition of metadata describing a new resource to the collection should not result in the metadata describing other resources in the collection to become incorrect.

We will call the need to satisfy these three conditions the 3C rule. The 3C rule tells us that to achieve good performance in search and discovery systems, the collection must be described with metadata satisfying the three metadata durability conditions. However, achieving compliance with them in practice might be problematic. It might require periodic updating of metadata describing the collection resources with new additions and changes to the underlying

metadata structures.

In terms of Type 1 metadata, it is necessary to ensure that all metadata fields describing a resource are populated, if that metadata field is applicable to the resource (completeness) and that this metadata information conforms to a single standard (consistency). For example, if all resources in the collection have an English title (even though the language of the resources might be different) then a new resource in French added to the collection should also have a title in English (completeness). Or if all resources are required to use a specific identifier, such as the *Document Object Identifier (DOI)*, in the identifier field, then all resources should have it (consistency).

In terms of Type 2 metadata, it is important to ensure, in addition, that the taxonomy (and its versions) used in the metadata creation process is applied as an annotation template consistently to all resources. For example, if a new set of concepts is added to the taxonomy and these concepts are used as metadata of a newly added resource, the remaining resources in the collection should be checked for their potential relevancy to these concepts. If the collection is large or the metadata creation process is subjective, then achieving all completeness, consistency and correctness for Type 2 metadata can be difficult.

In the case of Type 3 metadata, which have a relational nature, completeness becomes an issue. Adding a new resource with Type 3 metadata to the collection can cause incompleteness of some other Type 3 metadata describing

a different resource. This situation occurs depending on the properties of the Type 3 metadata relationship. For example, if the metadata describe a relationship that is symmetric, such as *similarity*, then by adding the metadata of a new resource to the collection, one is required to also modify the metadata of the existing resources (which are similar to this resource). The same holds when the metadata describe a relationship that is transitive, such as *follows*. In fact, correctness is another problem. By adding a new resource to the collections, relations between existing resources can be both created or disappear as a result of the change in the universe. For example, if the relation describes the closest resource (regardless of the definition of *closeness*), adding a new resource may invalidate the original relations. Consequently, the metadata should change.

If we define *metadata durability* as a quantity of how much the metadata describing resources in a collection remain consistent, complete and correct over time and as the collection grows, we can say that: The durability of Type 1 metadata is higher than that of Type 2 metadata which is higher than that of Type 3 metadata.

2.3.4 Discussion

Based on the equations, we now discuss the feasibility of manual annotation. It can be seen that when t_1 is small, providing type 1 metadata may be for

human annotators relatively effortless. Generating type 2 metadata may be still possible to perform in case t_2 and h are small or if the task is formulated as a *one-of* problem (exactly one taxonomy node is selected). However, specifying type 3 metadata can be performed by humans only for a very limited amount of resources. For example, if we assume that accessing and understanding a resource takes one minute, the annotation of a collection of 100 resources can take up to 165 hours. Furthermore, binary linking of resources is the most difficult metadata type to maintain as adding a new resource to a collection would have typically much higher frequency than changing a classification taxonomy or a set of possible keywords. Last, in certain collection types, such as multilingual collections, it may be for humans extremely difficult to perform such task.

To conclude, human performed link discovery does not scale up and can become infeasible even for very small collections. This makes link generation also theoretically unsuitable for collaborative approaches[4] which can be well applied to type 1 and type 2 metadata. A predominant approach is to generate links based on text analysis of documents or their type 1 or type 2 metadata. Current computer systems are capable of generating semantic similarity links in repositories containing up to one million of resources (Manning et al., 2008) when all possible pairs are checked, thus by algorithms with $O(n^2)$ complexity.

[4]This may be possible only for resources that have a high visit frequency by domain experts

For larger repositories, approximations calculated by algorithms with lower complexity can be used (see (Manning et al., 2008)). As humans are unable to carry out the link generation task, even algorithms with a relatively low precision and recall, in comparison to Type 1 and Type 2 metadata generation methods, are of real value.

2.4 Current approaches to metadata generation

There exist a wide range of approaches to metadata creation. These approaches differ in their suitability for producing metadata in the context of very large collections. They differ mainly in terms of the time needed to generate metadata and the degree the output satisfies the conditions of consistency, completeness and correctness described in Section 2.3.

A traditional approach to metadata creation is that somebody, for example a librarian, enriches a resource with all necessary metadata (resource cataloguing). Depending on the size of the collection and type of metadata, the time cost needed to fulfil this task individually ranges from fairly expensive to infeasible. The resulting metadata are likely to be quite consistent (as only one person performed the cataloguing), but they are also highly probable to be incomplete and often incorrect, because one person cannot be an expert on

all resources in a large collection. This approach is also nearly impossible to apply in multilingual collections.

Another commonly used approach is crowd-sourcing metadata from either authors or users. The advantage of the former is that authors are experts in the domain and this approach should therefore more likely lead to metadata with high correctness. The disadvantage is that the focus of authors can be too specific and this can therefore lead to the creation of inconsistent metadata. The advantage of the latter is that users can help with the maintenance/updating of metadata over time. However, as they are not necessarily the authoritative experts, their correctness might not be that high. Both approaches can be combined. The crowd-sourcing approach scales up relatively well for Type 1 metadata and Type 2 metadata (in collections with a large user community), but is insufficient for Type 3 metadata, where the growth in the size of the community cannot keep up with the growth in the amount of resources and their combinations (Knoth and Zdrahal, 2011).

The line of research called *metadata generation* or *automatic metadata generation* aims at providing tools that simplify the metadata creation process, improve the metadata generated by humans or extract metadata automatically from content. Table 2.1 provides some rough guidelines about the expected properties of certain approaches. The process of fully automatic metadata extraction has an advantage in consistency, completeness and speed. The success

type of metadata generation	feasibility (max collection size)	consistency	completeness	correctness
manually created by a curator(s)				
Type 1	Medium	Medium-High	Medium-High	High
Type 2	Small-Medium	Medium	Medium	Medium
Type 3	Small	Low	Low	Low
crowd-sourced				
Type 1	Large	Medium	Medium-High	Medium-High
Type 2	Large	Low-Medium	Low-Medium	Medium
Type 3	Medium	Low	Low	Low
extracted from content				
Type 1	Large	High	High	Medium-High
Type 2	Large	High	High	Medium-High
Type 3	Large	High	High	Medium

Table 2.1: The different approaches to metadata generation and their limits with respect to collection size and metadata durability characteristics - consistency, completeness and correctness, as defined in Section 2.3.3.

of state-of-the-art metadata extraction methods in terms of correctness ranges from quite poor to very good depending on the metadata field in question, the used data and the domain. However, the potential for achieving good correctness (precision) and completeness (recall) is high.

Greenberg et al. (2006) carried out a study in which she surveyed 216 experts about the use of automatic metadata generation tools. Participants were asked for their opinions on the feasibility and usefulness of automatic metadata generation for different Dublin Core metadata fields in terms of their accuracy and a suitable metadata generation level (manual, semi-automatic, fully automatic). The results (see Figure 2.4) indicated that the surveyed

experts anticipated greater accuracy with automatic techniques for metadata, such as ID, language and format (Type 1 metadata) than for metadata fields that require intellectual discretion, such as subject (Type 2 metadata) and description. The greatest scepticism has been reported on the `dc:relation` metadata field (Type 3 Metadata). While it is encouraging to see that the survey participants realised the potential of automatic techniques in achieving moderate to high accuracy, it is interesting to see that only low to moderate accuracy has been expected for Type 3 metadata. This is understandable due to the difficulty of the task and consequently we can expect this level of accuracy even if the task is carried out manually. Although the majority of participants still indicated that semi- or fully automated techniques would be useful for the `dc:relation` element (see Figure 2.5), it is also the element with the highest number of participants choosing fully manual approach as the most appropriate. We think this is due to the participants not being fully-aware of the limits of manual solution to this problem and the previously described metadata durability implications of the manual approach. We will show, in Chapter 6 that automatic generation of links matches the accuracy of humans performing the same task in large collections.

According to Greenberg (2004), there are two approaches to automatic metadata generation: *metadata extraction* and *metadata harvesting*. Metadata extraction is defined as an approach which automatically extracts meta-

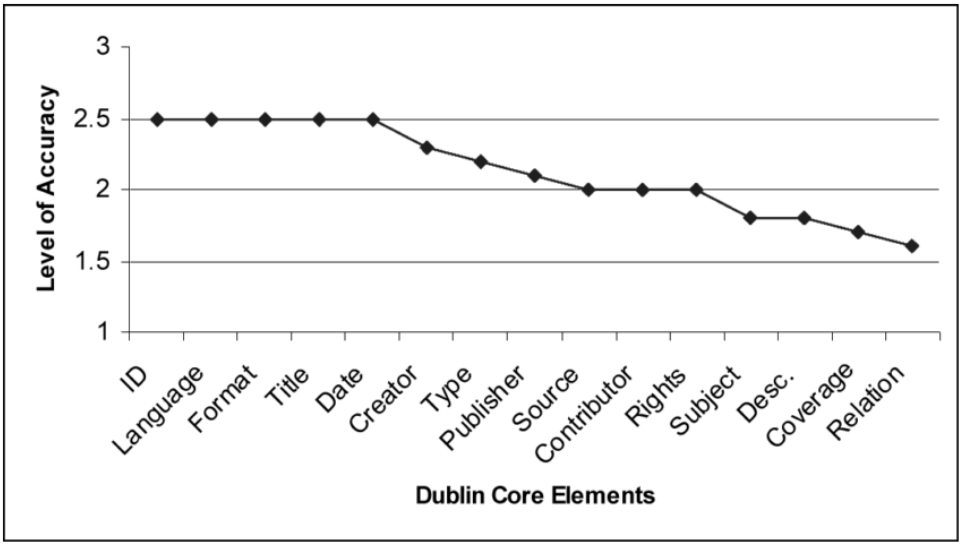


Figure 2.4: Expected accuracy for automatic generation of Dublin Core metadata elements according to the survey reported in (Greenberg et al., 2006). '3' meaning "very accurate", '2' meaning "moderately accurate" and '1' meaning "not very accurate".

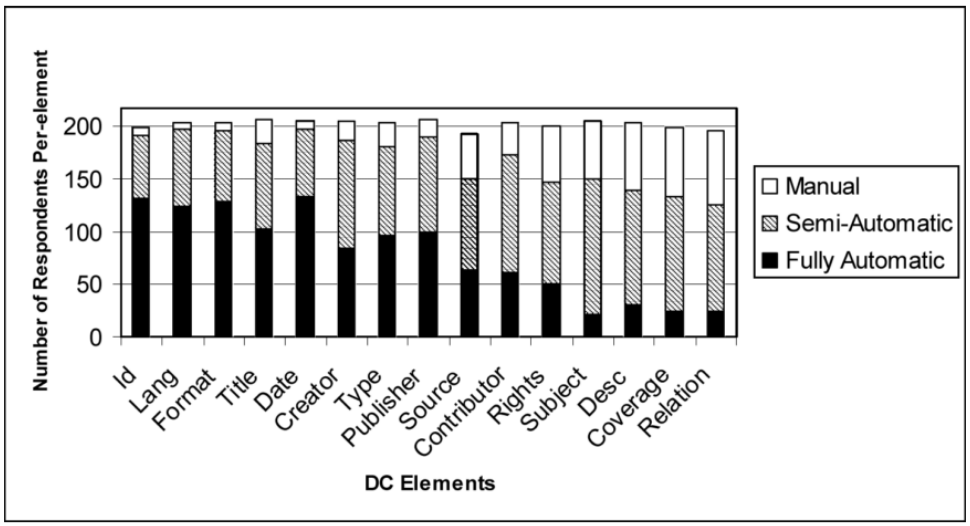


Figure 2.5: Appropriate metadata generation levels for Dublin Core according to the survey reported in (Greenberg et al., 2006).

data from the resource's content, while metadata harvesting is defined as a process which only collects metadata from existing structures, such as META tags in a Web page. However, in our opinion, to extract metadata only from the resource's content is in many cases not possible. While this might apply to Type 1 metadata, for automatic generation of Type 2 and Type 3 metadata, the access to the resource's content as well as access to external resources is essential. In addition, in the context of this document, metadata harvesting is not considered a metadata generation technique, because we see metadata generation as a process which generates some structured explicit information from content that contains the information in an implicit and often unstructured form.

There exist numerous approaches to the automatic extraction of different types of metadata many of which make use of some form of Natural Language Processing (NLP). These approaches have been in most cases designed to solve individual problems and they are typically not integrated to form a specific metadata generation package.

A typical example of a Type 1 metadata generation problem is keyword extraction. We have comparatively evaluated a set of existing automatic term recognition (ATR) methods in (Knoth et al., 2009). These methods are based on statistical characteristics of noun-phrases. These characteristics can draw on both the statistical information derived from the domain specific document

itself as well as information from some general purpose background corpus. The results of our evaluation indicated that there is a significant qualitative difference in the type of terms different methods automatically generate, though performance of these methods is typically fairly high. Other innovative work in Type 1 metadata extraction includes extraction of basic metadata fields (Kern et al., 2012), index terms (Erbs et al., 2013), indicator phrases (Daniel, 2012) and citations contexts (Bertin and Atanassova, 2012).

Generation of Type 2 metadata is usually approached as a text-classification problem. Text classification has been applied to many real world problems, such as spam detection. Its importance grew quickly with the amount of information available on the web. There exists a great variety of methods for text classification. Many machine learning techniques were explored in the context of text classification including naïve Bayes, decision trees, k-Nearest Neighbours, Rochio, neural networks and support vector machines (SVMs). An overview of the text classification methods can be found in (Manning et al., 2008; Sebastiani, 2002). One of the predominant approaches today is to convert documents into Vector Space Model (VSM) representation and to train machine learning classifiers on a set of labeled documents. The trained methods are then applied to an unseen set of documents. One of the common shortcomings of text classification methods is that they may require a great deal of supervision in terms of the amount of labeled documents. Techniques

that are able to learn and generalize from a very small set (*seed*) of labeled documents by relying on pattern acquisition from an unlabeled set of documents are becoming very popular. These techniques are often referred to as *weakly* or *semi-supervised* learning. One of the well known examples of weakly supervised learning is *active learning*. This is a technique that in each step determines an example, from an unlabeled set, the labeling of which is the most likely to help improve the performance of the algorithm and asks a user to provide this label. In this way, active learning minimises the size of the training set.

In the context of metadata generation and information discovery, hierarchical classification is particularly interesting. Large web directories available on the Internet, such as Open Directory Project or DMOZ, require high manual maintenance and can largely benefit from automatic approaches. Related work on hierarchical classification of content into conceptual hierarchies can be found in (Cesa-Bianchi et al., 2006; Dekel et al., 2004; Frommholz, 2001). For example, experiments with SVM applied to hierarchical classification of web content were reported in (Frommholz, 2001). Cesa-Bianchi et al. (2006) present an approach using a combination of the naïve Bayes algorithm and hierarchical SVMs to achieving good performance on test data.

Finally, generation of Type 3 metadata is the problem of automatic link discovery (Wilkinson and Smeaton, 1999), which is the focus of this thesis.

Although there is a great need for link discovery methods, the field is in comparison to Type 1 and Type 2 metadata generation relatively unexplored. The state-of-the-art in link discovery will be presented in Chapter 3.

2.5 Relationship types

The notion of using links to facilitate the exploration and navigation over resources is relatively old. In 1945, Vannevar Bush published an influential article (Bush, 1945) where he considered a future device called "*memex*." According to Bush, memex allows an individual to store all their books, records and communications. Bush then goes in his ideas further, by establishing links as an essential part of memex and claiming that they correspond to a natural way how our mind operates:

> "The human mind operates by association. With one item in grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of brain." — (Bush, 1945)

In this way, Bush predicted the emergence of the hypertext where resources are linked. An important extension of Bush's work has been developed in 1983 by Randall Trigg who realised that links are not semantically equal. Trigg has developed the first hypertext taxonomy of link types (Trigg, 1983).

His link types carry quite rich semantic information. For example, he envisaged links, such as *explanation*, *simplification/complication*, *continuation*, *critics*, *supported*. Although Trigg's work has been neglected by the specification of Hypertext Markup Language (HTML), which only support the `rel` attribute for some basic structural links (such as to mark the next page), additional more specialised link type taxonomies have been developed (Buckingham Shum et al., 2000; Mancini, 2005; Radev et al., 2008). Such taxonomies can be very useful for navigation and semantic representation of arguments (Uren et al., 2003). This is important, for example, for developing technology that can assist in the discovery of new arguments and counter-arguments. Such technologies have a direct impact on improving exploration in the digital space through effective use of metadata.

While both Bush and Trigg expected links to be built manually by users, we have made a case in this chapter demonstrating the need for more automatic approaches. An influential article on automatic identification of typed links based on the textual content of resources has been published by James Allan in (Allan, 1996) and also in his dissertation (Allan, 1995). Allan first classifies links into three categories:

1. *Pattern-matching links* — Can be discovered automatically using simple techniques (for example, using regular expressions/patterns).

2. *Manual links* — Cannot be discovered automatically using current tech-

nology. The opposite of the pattern-matching links.

3. *Automatic links* — Cannot be discovered in a trivial way, but can be recognised using statistical techniques.

Allan then focuses on the automatic links. This category of links include, for example, a relationship in which one document expands a topic discussed in the other document or a so called "*tangent*" relationship in which the same topic is discussed from two different perspectives. The main contribution of Allan lies in the development of a few heuristics that can be used for the discovery and typing of links and the algorithm, which can be used for the calculation. Allan's methods for automatic link typing represent documents using VSM. All documents are analysed by splitting them into smaller parts, such as paragraphs or for example, using topic segmentation techniques (Reynar, 1998). Similarity measures are then applied to all possible document pairs to recognise semantically similar segments and to generate links among them. This information is passed to a merging algorithm, which is used to consolidate links into a more simplified structure. Various hypotheses can be then applied to detect link types based on the pattern and the mutual position of the links.

While Allan's research is perhaps one of the most important pieces of work performed in link typing so far, there is no extensive evaluation of his approaches and it is also important to note that his automatic link typing meth-

ods cover only a very restricted subset of the link types originally presented by Trigg. An essential next step would therefore be to create a dataset that would allow the evaluation of such methods. This is basically a pre-requisite for any technological progress. We address this issue in Chapter 5.

2.6 Benefits and drawbacks of metadata standardisation efforts

With the evolution of the Web, there is an ever increasing need for services being interoperable, i.e. being able to communicate. Such interoperability can be achieved through adoption of common standards defining the communicated data structures and the communication protocols. In the context of large document collections, the standards typically apply to metadata schemas or ontologies describing documents (or their parts) and protocols to exchange them. An interesting aspect of the widely adopted metadata schemas, such as Dublin Core (DC), IEEE LOM or Europeana Data Model (EDM), is that they typically contain metadata fields of all the three metadata types discussed above. This makes it difficult for resources, described by metadata according to these standards, not to deteriorate in terms of their accessibility over time as discussed above.

A useful demonstration of the problem can be seen in the domain of open

access research outputs stored across the network of institutional repositories, subject-based repositories and open journal systems (later just repositories). These repositories widely adopted the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which defines a protocol for aggregating metadata of research outputs, described in the vast majority of cases in DC. The aim of repositories for adopting this protocol was to achieve interoperability. As the DC metadata are created in a distributed environment and in parallel at many organisations, they are, as a matter of fact, inconsistent, incomplete and incorrect (see Section 2.3.3). This makes it very complicated to develop systems that would provide a harmonised access to the whole dataset. Yet, this is vital for being able to provide exploratory search and discovery on top of these data. Certain guidelines for DC have been developed, such as the RIOXX Application Profile and OpenAIRE guidelines, trying to establish some good (but unfortunately mutually conflicting) practices that should lead to increased consistency, completeness and correctness. However, these guidelines are primarily addressing only Type 1 metadata, suggesting a certain level of helplessness to deal with Type 2 and Type 3 metadata interoperability through manual creation of metadata at this scale. We will present a system that generates Type 3 metadata from this dataset automatically in Chapter 8, providing good levels of completeness, consistency and correctness.

An essential question one might ask is about the relationship of meta-

data standards and automatic metadata generation methods. Historically, metadata (used widely in libraries before the information era) were created manually. Today, the work on new metadata schemas and classification systems/ontologies is typically still driven by a similar process as in the past. This process asks, primarily, what needs to be represented, not taking into account whether the metadata will be produced manually or automatically. It is questionable, whether such approach is actually logical. The evidence shows that complicated subject classification systems or link taxonomies/ontologies, including those developed in (Buckingham Shum et al., 2000; Mancini, 2005; Radev et al., 2008; Trigg, 1983) have not yet become widely adopted, despite their obvious potential benefits. We believe that a likely explanation is that they were never supported by a reliable suite of automatic metadata generation tools that would make their use scalable. While it is reasonable to create metadata standards that should motivate the development of new or more advanced metadata generation technology, it needs to be understood that there might be a trade-off between the expressivity of the standard and its potential for wide adoption in practice. This can be, for example, due to the issue of scalability of manual metadata provision or yet non-existent technology for automatic metadata generation. We argue that when manual metadata provision does not scale, metadata standards should be designed with the awareness of potential benefits and also limits of automatic metadata generation tools.

2.7 Objectives of the thesis

This chapter has so far motivated and discussed the research problem and has introduced the state-of-the-art in the related areas. This forms the foundation for defining the gap and subsequently specifying the objectives of the thesis, formulating the research questions to be answered and describing the research methodology.

2.7.1 Defining the gap

At the beginning of this chapter, we motivated the research problem. We discussed that the thesis focuses on the specific issues related to the problem of efficiently and effectively supporting the discovery of resources through the automatic identification of links. We mentioned that discovery can be seen as an important part of exploratory search, which is currently a growing area of information retrieval.

Later on, we discussed the importance of good quality metadata to support discovery. We identified the main types of metadata and analysed their differences. We learned that relational metadata (Type 3 metadata) are difficult to acquire and maintain, while at the same time particularly important for discovery. Despite the importance of the link discovery research problem, which essentially tries to automatically generate relational metadata, there are

still many research opportunities. There is a lack of methods and comparative studies. Datasets that could be used for evaluation of these technologies are difficult to obtain. In comparison to some of the traditional information retrieval tasks, the progress of which is regularly evaluated at evaluation forums, such as TREC, the link discovery field is still fairly young. In the recent years, the main evaluation forums for link discovery technologies have been INEX and NTCIR CrossLink. The aim of these forums has been to encourage the continuous development of these systems, improve and compare methods, understand evaluation metrics and to develop methods that are capable of linking resources across language boundaries.

We then discussed the various semantic types of information relational metadata can carry. We noticed that the problem of automatically detecting the semantic type of a relationship (link typing) is an area where sufficient research has not yet been carried out. While it seems that there is no doubt that link typing is important, it is not clear how it could be performed automatically in a general domain and based on which characteristics. There is also no consensus in terms of which semantic relationships should be distinguished.

Finally, we mentioned common approaches and standards using which relational metadata are currently being expressed and discussed the benefits and drawbacks of standardisation efforts in this area. We learned that the most widely used approaches and standards for describing resources are fairly weak

in their ability to carry information about semantic relationships. However, it also seems that standardisation efforts that would be targeted at describing relationship types can only be successful if they follow an approach in which they realistically take into account the possibilities of current technologies in terms of distinguishing various semantic types.

2.7.2 Research questions

In order to address the current issues and challenges of discovery, described in Section 2.7.1, we formulated the following central research question:

CQ: How to efficiently and effectively support the process of identifying links between semantically related resources in large textual collections and use them to facilitate discovery?

After carrying out the literature review (Chapter 3), we identified the following set of more detailed sub-research questions and goals to focus on:

RQ 1: Is it possible to identify properties that would suggest which pairs of textual resources are more likely to be linked by people?

RQ 2: Can some of the properties identified by answering RQ 1 be used to determine the semantic type of a link?

RQ 3: How can we detect links between textual resources written in different languages?

RQ 4: How shall we interpret the performance achieved by link discovery methods and how does the technology compare to the ability of humans to carry out the same task?

To substantiate the research work, our goals are to:

Goal 1: Design new link discovery methods and evaluate them under the umbrella of an international evaluation conference, such as NTCIR, in direct competition with other research teams.

Goal 2: Show that link discovery techniques can be deployed in large document collections to facilitate access to public knowledge.

2.8 Methodology

We start our research by reviewing the state-of-the-art in the area of link discovery (Chapter 3). While the goal of Chapter 2 was to review work in related areas and by doing so motivate the selection of the central research question, the state-of-the-art chapter focuses on reviewing the work directly associated with the research questions. It is therefore more specific in terms of methods and the target domain. Chapter 3 is fundamental as it provides the basis for pursuing the research work necessary to answer the research questions and achieve the goals listed in Section 2.7.2. It also provides further motivation justifying their selection.

Our approach is then to take the research sub-questions, listed in Section 2.7.2, one by one and dedicate a chapter of the thesis to answering each (or a few) of them. To ensure that the thesis contains a significant amount of material worthy of publication, which is specified as a criterion for the award of the PhD degree by the Open University, our approach for the creation of each of these chapters (Chapters 4-8) follows three steps. Firstly, we carry out the research addressing the research question(s). Secondly, the research and its outputs are documented by writing a chapter of the thesis. Thirdly, we publish the thesis chapter as a conference or a journal paper typically slightly modifying the chapter's narrative, as needed. This means that all research presented in Chapters 4-8 has already been published and, in the case of conference papers, also presented. Following a peer-review process, we have later also adapted the thesis chapters to integrate useful suggestions provided by the reviewers. Finally, our approach also follows the research strategy that the outcomes of each chapter can influence the research direction. Therefore, the research was also, where possible, pursued largely in the order in which the research sub-questions are listed.

To further substantiate the research work described in the thesis, enable a comparative evaluation and provide alignment with relevant current initiatives, we also report on (a) our participation in two consecutive link discovery evaluations organised by NTCIR CrossLink in 2011 and 2013 and (b) present

our work on the development of a large scale aggregation system CORE (COnnecting REpositories), which uses link discovery technology to interlink millions of open access research papers, with the aim to improve access to public knowledge.

2.9 The structure of the thesis

Chapter 3 presents a critical review of automatic link discovery approaches. It formally defines the link discovery task, classifies existing link discovery approaches reported in the literature according to granularity, type of input and application domain, and reviews them. The chapter presents evaluation metrics used later in the thesis and introduces the main challenges in the evaluation of link discovery systems. Overall, the chapter provides the foundations needed in the subsequent chapters of the thesis.

Chapter 4 investigates the behaviour of people in linking content and analyses its relationship to semantic similarity. The chapter addresses RQ 1. Its content was later published in the following paper:

Knoth, P., Novotny, J. and Zdrahal, Z. (2010) Automatic generation of inter-passage links based on semantic similarity, The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China

Chapter 5 is motivated by the results of Chapter 4. It explores if the value of semantic similarity can be used to predict the relationship type between two

textual fragments. The chapter addresses RQ 2 and resulted in the following publication:

Knoth, P. and Zdrahal, Z. (2011) Mining Cross-document Relationships from Text, The First International Conference on Advances in Information Mining and Management (IMMM 2011), Barcelona, Spain

Chapter 6 designs and evaluates a set of new methods for document-to-document cross-language link discovery (CLLD). Importantly, using the assumption that links connecting pieces of information have a semantic foundation, we make use of the multi-lingual environment to study their disparity in different languages. This enables us to assess the subjectivity of the linking task and better understand the practical performance limitations of automatic link discovery systems. The chapter addresses RQ 3 and RQ 4 and was published as the following paper:

Knoth, P., Zilka, L. and Zdrahal, Z. (2011) Using Explicit Semantic Analysis for Cross-Lingual Link Discovery, Workshop: 5^{th} International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011), Chiang Mai, Thailand

Chapter 7 presents new noun phrase-to-document CLLD methods we submitted to two evaluation forums (NTCIR-9 CrossLink and NTCIR-10 CrossLink

2). In addition to the design of the methods and their evaluation, we discuss the differences in our approaches to CLLD at the two consecutive evaluation conferences, explain how they relate to each other and compare them with those of other evaluation participants. Last but not least, we identify and debate various issues related to the evaluation methodology used at CrossLink and suggest improvements. The chapter further deals with RQ 3 and RQ 4 and addresses Goal 1. The research work reported in the chapter resulted in two papers:

Knoth, P., Zilka, L. and Zdrahal, Z. (2011) KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis, NTCIR-9: Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Tokyo, Japan, pp. 495-502

Knoth, P. and Herrmannova, D. (2013) Simple Yet Effective Methods for Cross-Lingual Link Discovery (CLLD) - KMI @ NTCIR-10 CrossLink-2, NTCIR-10 Evaluation of Information Access Technologies, Tokyo, Japan

Chapter 8 introduces the issues related to (mainly programmable) access to research publications. This access is essential for realising the potential of link discovery in a domain where exploratory search and discovery experiences are needed. The inability to apply this technology in this domain led us to the development of the CORE aggregation system, which we also integrated with

link discovery software. We also report on the wider benefits of the effort to bring these data together and present how we perform link discovery in CORE. The chapter addresses Goal 2 and resulted in the following publication:

Knoth, P. and Zdrahal, Z. (2012) CORE: Three Access Levels to Underpin Open Access, D-Lib Magazine, 18, 11/12, Corporation for National Research Initiatives

Finally, Chapter 9 summarises the overall original contribution of the thesis, discusses the limitations, future work and concludes.

2.10 Conventions

There are numerous, sometimes conflicting, suggestions about the presentation format and the use of language in a PhD thesis. We would like to inform the reader about the conventions we have selected for this thesis, as this might ease the interpretation of the text and improve the reading experience.

Each of the five chapters that follow the literature review (Chapters 4-8) first introduces the research question(s) and goals it addresses and provides a summary of contribution with respect to them at the end of the chapter. Chapter 9 then summarises the overall contribution of the thesis only with respect to the central research question.

The thesis uses italics for three main purposes. For the introduction of new technical terms and phrases, for titles used in bullet points or numbered lists

and in specific reference to the wording of the research questions.

We have also made an informed choice to use "we" consistently throughout the thesis. While some argue the use of "we" in academic writing sounds pretentious (Robinson, 2014), others claim the use of "I" sounds egoistic instead (*Use of I, we and the Passive Voice in a Scientific Thesis*, 2014). Reviewing a few past PhD dissertations submitted and successfully defended at the Open University suggests that both approaches are possible. Despite the use of "we" in sections discussing the research contribution, we always refer to the original research contribution of the thesis author unless explicitly stated otherwise.

Chapter 3

State of the Art in Cross-Document Link Discovery

In the previous chapter, we discussed the infeasibility of manual creation and maintenance of links between resources in digital repositories. This creates an acute need for automatic or semi-automatic approaches to link discovery.

In this chapter, we provide the state of the art in automatic link discovery and related areas. We start by defining the link discovery task (Section 3.1). We then review existing work in the relevant areas (Section 3.2). Finally, we discuss approaches to evaluation of link discovery systems (Section 3.3).

3.1 Task definition

Throughout the thesis, we will use the following definition of link discovery:

The automatic link discovery task can be defined as follows: Let D_s and D_t

be collections of documents $d_s \in D_s$ and $d_t \in D_t$ respectively. Let $d'_s \in D'_s$ and $d'_t \in D'_t$ denote the set of all substrings of $d_s \in D_s$ and $d_t \in D_t$ respectively. Let S and T, denoting the sets of link sources and targets, be defined as $S = \{s|s \subseteq d'_s \in D'_s\}$ and $T = \{t|t \subseteq d'_t \in D'_t\}$. For example, s and t can be strings corresponding to the text of the whole documents, paragraphs, sentences, noun phrases or words appearing in $d_s \in D_s$ and $d_t \in D_t$.

The goal of link discovery is to find a binary relation $\rho \subseteq S \times T$ defined in terms of pairs $\langle s_i, t_j \rangle$ such that the pairs are interpreted by a human evaluator as carrying the same semantic relationship. For example, ρ can be interpreted as *is similar* or *is related*, or even as more specific relationships, such as *is_the_same*, *expands* or *contradicts*.

If $\forall \langle s_i, t_j \rangle \in \rho, s_i \in d_k \Rightarrow t_j \in d_k$, we talk about *intra-document* link discovery. On the other hand, if $\forall \langle s_i, t_j \rangle \in \rho, s_i \in d_k \Rightarrow t_j \in d_l, k \neq l$, we talk about *inter-document* or *cross-document* link discovery. Following the goals of the thesis, we will focus our attention on cross-document link discovery. If $D_s \neq D_t$, we call the task *cross-collection* link discovery.

If D_t was representing a set of any resources, not just documents (as in our definition), we would extend the problem space to ontology-based link discovery/population (Cimiano, 2006), which tries to link (textual) entities appearing in documents/web resources to concepts in an ontology. Please note the term link discovery is also used by the semantic web (linked data)

community to refer to the problem of explicitly establishing RDF links between entities across data sources (Volz et al., 2009). Although research in these areas is slightly related to some of the research questions addressed in this thesis, it is not our primary focus.

The ρ relation must satisfy the usual properties implied by semantics attributed to its name, such as *is the same* is symmetric, transitive and reflexive, *is_similar* is not transitive, *expands* is antisymmetric. In practice, link discovery tasks are typically concerned with the more general relationships, such as *is_related*, dealing with the problem of how to discover and identify pairs of related resources in a large collection. The problem of automatically finding more specific relationships, such as *expands*, is then reduced to the filtering of the set of more general relationships acquired using link discovery methods. We call the line of research dealing with this problem *link typing*.

3.2 Review of link discovery methods

Link discovery methods can be divided according to the following criteria:

- The granularity of the string, such as document, paragraph, noun phrase or word, used as the link source and target.
- The type of the input information based on which links are discovered.
- The use cases in which the method is applied.

We will now discuss the different classes of link discovery systems and review the existing approaches to link discovery.

3.2.1 Link discovery methods according to granularity

Link discovery approaches working at different levels of link source and target granularity are suitable in different contexts. The link source and target granularity is typically at the level of:

- document
- paragraph
- segment (a semantically self-contained segment spanning one or multiple sentences)
- sentence
- noun phrase (named entity or concept)
- word

Different combinations of granularity of the link source and target are possible. For example, a *paragraph-to-document* link can refer to linking a quotation to its document of origin, *noun phrase-to-noun phrase* can be used to link a product to a company, a *noun phrase-to-document* link can connect a

concept to a document with its definition and *paragraph-to-paragraph* link can associate a passage to its more detailed version.

3.2.1.1 Document-to-document link discovery

Document-to-document[1] link discovery is probably the most widespread type of link discovery, which can often also be seen as a form of content recommendation. In terms of the task definition, it is closely related to traditional information retrieval (IR) with the query being the whole document (*query by example* or *more like this*).

Consequently, the task is nowadays usually approached using IR solutions. Typical approaches find semantically related documents by calculating their semantic similarity based on term-document vectors (Allan, 1997; Green, 1998; Zeng and Bloniarz, 2004). The term-document vectors are usually created by processing the text of the resources applying techniques, such as tokenization, stop words filtering, stemming, weighting (e.g. *tfidf*), and normalisation. More advanced approaches perform additional projections or reductions of the term-document vectors, such as Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) or Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas and Harshman, 1990). A range of semantic similarity measures can be then applied to the calculation of similarity between the term-document vectors of two resources. Cosine, overlap, dice and Jaccard coeffi-

[1] sometimes referred to also as *file-to-file*

cients are widely used measures for the calculation of similarity $sim(\overrightarrow{x}, \overrightarrow{y})$ of the document vectors $\overrightarrow{x}$ and $\overrightarrow{y}$. If the calculated similarity is higher than a given threshold τ, then a new link is generated.

An early position paper/survey on automatic link discovery systems has been published by (Wilkinson and Smeaton, 1999), considering mostly the document-to-document scenario in the content recommendation context. This paper clearly articulated the need for more discovery experiences as a result of the large amount of information available online (and the difficulty of manually interlinking this information). Wilkinson and Smeaton (1999) also mention the issue of automatic identification of link types, which is missing on the current Web. Since that time, many link discovery systems have appeared. More recently, link discovery systems have been used in large digital repositories, such as PubMed[2] or ACM Digital Library[3] and academic search engines, such as Google Scholar.

Although methods measuring semantic similarity are widely used in practice in this context, more work is still needed to understand their application in link discovery. For example, it is not clear whether high similarity is a good predictor for generating a link as assumed by Wilkinson and Smeaton (1999) or what is the qualitative impact of document length on automatic link discovery (irrespective of the use of length normalisation techniques).[4] Green

[2] `http://www.ncbi.nlm.nih.gov/pubmed`
[3] `http://dl.acm.org/`
[4] We investigate these issues in Chapter 4

(1999) studied how the quality of the generated links is influenced by the use of background knowledge. More precisely, Green investigated whether links discovered based on ontologies, using a lexical chaining method, are better than links discovered based on simple term-repetition document vectors. It appeared that his results were not statistically significant and so it was not possible to support his hypothesis.

The bottleneck of all link discovery approaches based on pair-wise semantic similarity is in the high number of term-document vector pairs for which similarity needs to be calculated. Elsayed et al. (2008) developed a distributed algorithm for calculating semantic similarity using the MapReduce paradigm[5]. The method exhibits linear growth in running time and space with respect to the number of documents in the collection, as tested on a 900k large newspaper corpus. In fact, the crucial optimisation is not in the parallelisation of the similarity calculation algorithm, but mainly in reducing the number of term-document pairs by eliminating a certain percentage (for example, 1%) of the most commonly appearing words from the term-document vector (see Figure 3.1). This, in turn, dramatically reduces the number of pair-wise comparisons (as document pairs not sharing any term are, thanks to the inverted index structure, easily removed from consideration). The *df-cut* is thus an efficiency rather than an effectiveness motivated technique.[6]

[5]http://en.wikipedia.org/wiki/MapReduce

[6]A similar cut-off technique is applied in the CORE system, presented in Chapter 8. See Section 8.4.3 for details.

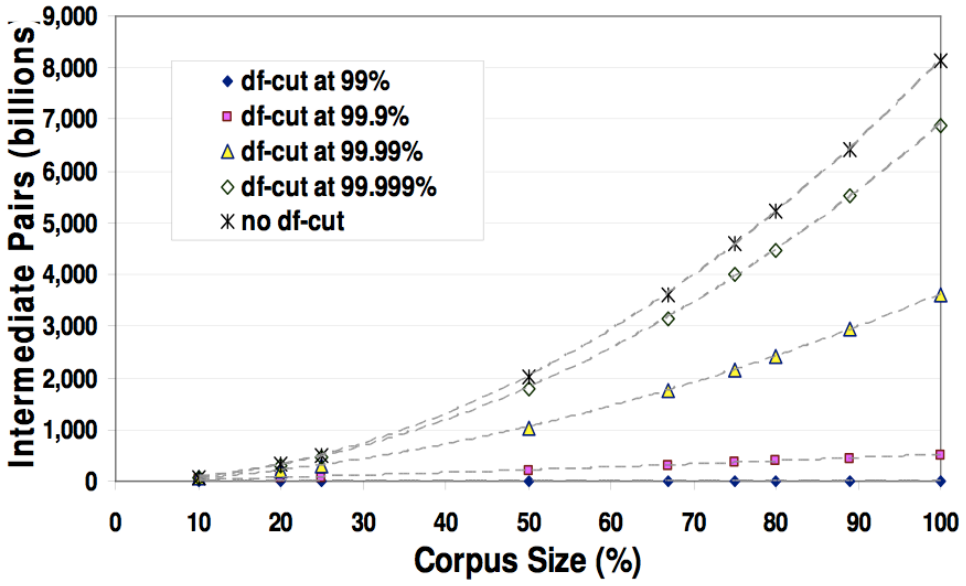


Figure 3.1: The effect of df-cut on reduction of pair-wise comparisons. Please note that the 99% df-cut curve effectively represents only a 1% cut. The original figure is taken from (Elsayed et al., 2008).

The current research in document-to-document link discovery addresses also the problem of automatic linking of information in heterogeneous and/or multilingual environment. The issue of linking news articles to blog entries discussing them has been addressed by Ikeda (2006) and the problem of discovering implicit links from online news to social media by Tsagkias et al. (2011). One of the interesting observations of this study is the relatively low level of vocabulary divergence between news articles and relevant mentions of these articles in certain social media (Twitter, Digg, blogs, Delicious), while there is a higher divergence between news (including their comments) and relevant Wikipedia articles (Figure 3.2). This vocabulary divergence between

some data sources might suggest that high similarity of documents might not always be a sufficient indicator for establishing a link.[7]

Linking semantically related documents across language barriers has to deal with a similar problem of vocabulary divergence. While mapping of documents from their lexical representation to some interlingual semantic representation can be established through dictionaries or natural language processing techniques, such as LDA or Cross-Language Explicit Semantic Analysis (CL-ESA) (Sorg and Cimiano, 2008*a*), a certain level of disparity is likely to stay due to cultural differences. Document-to-document cross-language link discovery has been, for example, investigated by Smet (2009) who tried to connect news articles in Dutch and English according to the events described in those articles using LDA modelling. Sorg and Cimiano (2008*b*) addressed the problem of automatically discovering missing cross-language links between corresponding Wikipedia articles.

3.2.1.2 Noun phrase-to-document link discovery

Noun phrase to document link discovery is directly related to the problem of automatically inducing hypertext structure in textual documents originally created without hypertext links. Typically, the problem is approached by first identifying suitable textual units that are good candidates for acting as hypertext links (anchors), often involving disambiguation of the anchor sense.

[7] We further investigate this issue in Chapter 4.

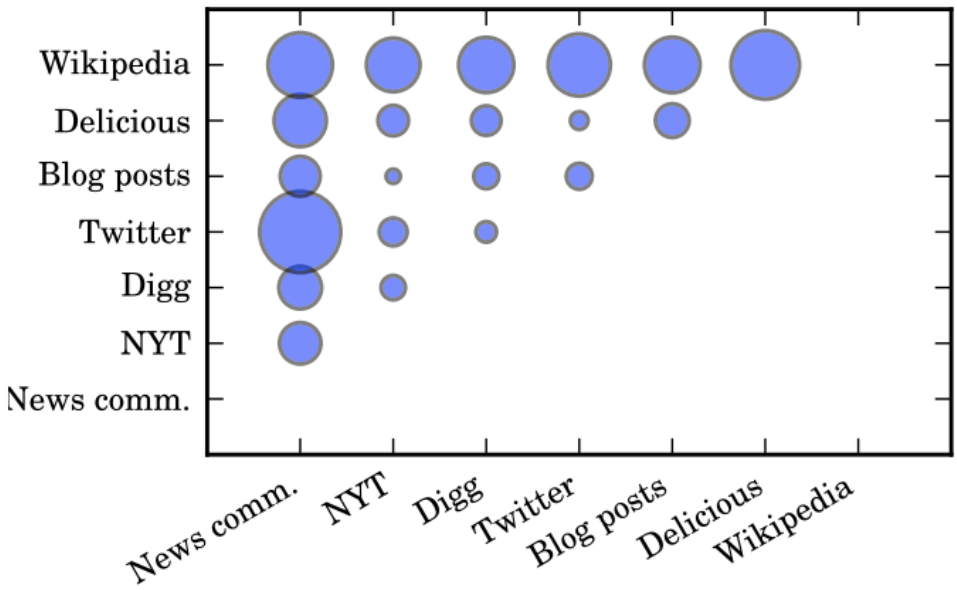


Figure 3.2: Average symmetric KL-divergence between New York Times articles and explicitly linked social media utterances from Digg, Twitter, blog posts, New York Times comments, Delicious and Wikipedia. Larger circles indicate a higher degree of divergence and hence a bigger difference in vocabulary. The figure is taken from (Tsagkias et al., 2011).

In the second step, appropriate documents (link targets) are detected.

A special but well known use of noun phrase-to-document link discovery systems is the task informally referred to as *wikification*. Given a document, an unlinked (orphan) wiki page or any other text, the goal is to enrich the document with links pointing to appropriate wiki pages. This improves the browsing experience of users in wikis, indicating to them what content is available and might be worth exploring. The link discovery system must generate a reasonable proportion of links with respect to the text length, in order not to overload the user, and the links should also point to the correct sense in which they are used.

Well known research studies on wikification include the *Wikify!* system developed by Mihalcea and Csomai (2007), which was to our knowledge the first attempt to use Wikipedia as a resource for link discovery. This approach has been later extended by Milne and Witten (2008) who used machine learning techniques to determine parameters of a simple sense disambiguation model based on relatedness and commonness of terms. Given the importance of this task and its relatively large application area, a forum for development and evaluation of these link discovery systems has been established at the Initiative for the Evaluation of XML retrieval (INEX), which is with TREC, CLEF and NTCIR one of the four main Information Retrieval Evaluation Forums. INEX: Link the Wiki track, took place in 2007, 2008, 2009 and 2010. The task has been then extended to a multilingual environment (English, Japanese, Chinese and Korean) and moved to the NII Testbeds and Community for Information Access Research (NTCIR) project. So far, there were two evaluations of this task at NTCIR-9: CrossLink (2011) and at NTCIR-10: CrossLink 2 (2013).[8] Dozens of systems and methods submitted by many international groups were comparatively evaluated at these forums.

While many research studies address the wikification problem in the context of Wikipedia, the methods are often also applicable to corporate wikis, knowledge bases or other similar systems. For example, Hoffart et al. (2009)

[8]We have participated to both NTCIR-9: CrossLink and NTCIR-10: CrossLink 2 achieving very good results. The developed link discovery methods are discussed in Chapter 7.

presents software tools for integration of NLP techniques with wikis, targeted especially at corporate environments. The tools provide a semi-automatic support for organising the content by suggesting relevant links between wiki pages.

This type of link discovery is also similar to the problem of knowledge acquisition and ontology population (Cimiano, 2006). A well-known forum for comparative evaluation of these systems is Text Analysis Conference (TAC) Knowledge Base Population Track, see for example (McNamee et al., 2009). However, since knowledge acquisition and ontology population does not primarily focus on improving exploratory and discovery experiences, it is out of scope of this thesis.

3.2.1.3 Other link discovery tasks

Link discovery tasks other than document-to-document and noun phrase-to-document are perhaps less well represented in the research literature. However, this does not make them in any way less valuable or important. As there is a number of combinations of the granularity of the source and the target that are possible, we will only discuss the more prominent examples. The common aim here is to identify, as the link target, a unit of a lower granularity than the whole document. These types of link discovery are therefore often aimed at improving the navigation experience in digital collections containing long

documents.

Discovering links pointing to units of a smaller granularity than a document can be seen as a task of *passage* or *focused* retrieval, a subdomain of information retrieval where the search engine locates the relevant information inside the document, instead of only providing a reference to the document. The standardised testing of search engines in this domain is performed by INEX. INEX played an essential role in formalising the link discovery task. It established the framework for the evaluation of systems generating links of a different granularity than *document-to-document* links. One of the INEX evaluation tracks is the Link-the-Wiki Track, which includes a tasks to analyse the text of a resource and to recommend a set of incoming and outgoing links from an anchor text to the Best Entry Point (BEP) in other documents in the collection. This means that the anchor text is linked to a specific position in the target document, i.e. the BEP to start reading the referenced material from. However, as the results of the INEX evaluation show (Huang et al., 2009; Trotman et al., 2009), performing and evaluating this task is complicated. This is due to the fact that in Wikipedia, the overwhelming majority of BEPs are located at the beginning of the article and the median reported BEP is only 311.5 characters from the start (Kamps et al., 2010). This means that a naive system that assigns all best entry points at the beginning of the document can perform well in this evaluation. This indicates the need for a

more robust evaluation framework for best entry points, perhaps one not using Wikipedia as a corpus.

An interesting algorithm for *segment-to-segment* link discovery is presented in (Kolak and Schilit, 2008). The authors developed a scalable method for mining repeated word sequences (quotations) from very large text collections, with the aim to improve user navigation. The method has been integrated with the Google Books archive and tested on a corpus of over 1 million books. The algorithm allowed users to navigate to popular passages across documents, which were not originally constructed as hypertext. An interesting issue in this task is the problem of identifying suitable units of granularity for both the link source and the target and tolerating a certain level of variation in quotation texts. The problem is approached using a technique called *shingling*. Shingles refer to k-consecutive tokens from a document. When a document is shingled, then all unique k-shingles are extracted from it (and they overlap as shingles on the roof). The exact matching of these shingles across documents can then be used to efficiently identify equivalent multi-word sequences across texts and determine appropriate boundaries for the quotations. The use of shingles with different value of k in conjunction with a ranking procedure allows for the detection of variations of the same quotation appearing in different documents. The evaluation of the approach indicated that over 88% of the discovered quotations were later confirmed by human readers, suggesting a fairly high

precision of this approach.

A very special yet highly interesting type of link discovery is its use for supporting the mining of research literature for new scientific discoveries. The idea here is to help identify non obvious (*hidden*) relationships between concepts (noun phrase-to-noun phrase link discovery). These relationships are not explicitly stated in the text, but are supported by the evidence present in the textual collection. This problematic was already discussed by Swanson (1986) in his paper *Undiscovered Public Knowledge* providing as examples of such scientific discoveries the relationship between magnesium deficiency and migraine or fish oil and Raynaud's disease. These Swanson's discoveries have been simulated by automated techniques by (Weeber et al., 2001).

The method is based on the premise that one publication may state the relationship between two concepts A and B, while another can report on the relationship between B and C. If no one has reported on the relationship between A and C, this association can be considered to be new. However, since the two pieces of information are not related directly, there is only a hidden connection (Weeber et al., 2001). The assumption used by automated discovery methods is that concept B connecting concepts A and C is likely to be a rare term, hence the connection is not obvious. The weakness of the well-known ARROWSMITH system following this assumption (Smalheiser and Swanson, 1998) was a large space of generated hypotheses, i.e. links, and consequently

the need of human expertise in their evaluation. The RaJoLink system (Petrič et al., 2009) addresses this problem by developing a semi-automated way of suggesting which relations might have more potential for new discoveries and are therefore good candidates for further investigations.

3.2.2 Link discovery methods according to the type of input data

The type of input data typically has a significant effect on the performance of link discovery methods. We can classify link discovery methods according to the type of input data into:

- *link-based*
- *based on semi-structured data*
- *purely text-based*
- *hybrid*

Link-based approaches exploit information about the structure of the existing link graph to find new (missing) links. Methods based on semi-structured data make use of information, such as document titles, headings and other forms of existing markup to generate links. Purely text-based approaches depend on information retrieval and natural language processing techniques to

discover new links. Finally, hybrid methods combine any of the previously mentioned approaches. Generally speaking, link-based methods have been shown to lead to very good results. Semi-structured methods can match the performance of link-based approaches on certain tasks. Existing purely text-based methods typically do not perform as well as the two above mentioned approaches, but they are most widely applicable and have probably the highest potential for improvement. Many of the currently applied link discovery methods combine several techniques (in order to exploit the available input information as much as possible), which puts them in the category of hybrid methods.

Overall, the variability of input data available in different link discovery tasks makes it complicated to compare systems unless they operate on the same dataset and have a shared goal. In this respect, data and results from evaluation conferences, such as INEX and NTCIR, are of vital importance, allowing us to draw conclusions about which approaches work better than others.

3.2.2.1 Link-based link discovery

Link-based link discovery methods discover new links by exploiting the patterns in the existing link graph. It has been demonstrated that information in the link graph is in many tasks very valuable and link-based approaches

can therefore achieve high performance. On the other hand, they cannot be applied unless a substantial part of the collection has been already interlinked. This causes them not to be applicable in the majority of link discovery tasks.

Itakura and Clarke (2008) developed a simple algorithm for the INEX:Link the Wiki Track, which first processes the entire collection of links in the Wikipedia corpus to obtain the most likely assignments of terms to the pages (concepts) they are linked to. They calculate a ratio they call γ:

$$\gamma = \frac{\text{number of pages that have a link from anchor } a \text{ to a file } d}{\text{number of pages in which } a \text{ appears at least once}} \tag{3.1}$$

Their algorithm only generates new links if γ is higher then a threshold (in their case 0.6). What is interesting about this approach is that despite its simplicity, it has been reported to work reasonably well in terms of precision/recall characteristics. A possible explanation to this is that the method chooses to generate only links, about which it is highly confident. The disadvantage of the method is that it will not generate a new link for highly ambiguous anchors (see more in Section 7.3). It chooses certainty over trying to disambiguate more complicated cases. The algorithm is also unable to generate new links for anchors that have not been linked somewhere in the corpus yet.

Jenkinson et al. (2008) worked further on investigating the features of this method and proposed some changes slightly improving the performance. How-

ever, the changes are cosmetic and refer only to the way anchors are pre-processed and the text is normalised. Another method fully relying on the link graph has been developed, for example, by (Lu et al., 2008).

An improvement to dealing with the disambiguation problem by exploiting the link graph has been proposed by (Milne and Witten, 2008). The idea is based on measuring the semantic similarity of two concepts representing Wikipedia articles by comparing their incoming and outgoing links. Formally, this is represented using a measure developed in (Milne and Witten, 2007):

$$relatedness(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|, |B|))} \tag{3.2}$$

where a and b are the two articles of interest, A and B are the sets of all articles that link to a and b respectively, and W is the set of all articles in Wikipedia. A problem of this measure is that it does not take into account the semantics of the words on the article page, relatedness of content is simplified to the relatedness of the link graph. Another problem of the measure is that it uses information, which is often not available at the time of calculation, i.e. we typically do not know (all or most of) the links that are incoming/outgoing to the article that is being enriched with links.

Consequently, different measures to tackle these problems based on textual and semi-structured information have been proposed in (Granitzer et al., 2008). Most of the successful approaches today, such as (Fahrni et al., 2011;

Knoth and Herrmannova, 2013), combine the benefits of the link-graph information with the semi-structured and textual context of the anchor's occurrence to perform more sophisticated sense disambiguation than the one reported in (Itakura and Clarke, 2008). The results published by Knoth and Herrmannova (2013), discussed in detail in Chapter 7, suggest that using the textual context in the disambiguation component can actually lead to better results than fully relying on the link graph.

3.2.2.2 Semi-structured link discovery

Semi-structured link discovery methods discover new links using semi-structured information. They are typically developed using document collection specific knowledge, for example, naming convention of document titles or correspondence of anchor text to document titles. Methods based on such knowledge can achieve high performance, while having fairly low computational requirements. The main disadvantage of these methods is that they are notoriously difficult to port to different collections or similar tasks.

The algorithm, which is probably the purest representative of this class, has been developed by Geva (2007) and also submitted to the INEX:Link the Wiki Track evaluation. It is based on the idea of putting all page titles in a hash map and then searching for those titles in the text we want to interlink. The algorithm also uses the assumption, which is based on the trivial observation

that longer phrases are more likely to be suitable link anchors than shorter phrases. Geva points out that his algorithm is computationally very light and it does not make any assumptions about the existing link structure.

However, this approach has also a number of drawbacks. First, it makes the assumption that anchors used as links in the collection should only point to pages where the title is equal to the anchor. This assumption is widely valid in wikis and encyclopaedias, but is not applicable on its own in collections like newspapers or research articles. Second, the method ignores the word sense disambiguation problem. The method was tested at the time English Wikipedia contained 660k articles. Today, it is more than six times larger. This means that there are also significantly more articles with the same title talking about different topics, than there were at that time. According to the description of the method, it is likely the performance would drop as the collections gets larger. On the other hand, there is no problem for new systems to employ, in addition to this method, a disambiguation component. Such a disambiguation component can actually use the information about the same titles to cut down the number of possible targets for an anchor (Fahrni et al., 2011; Granitzer et al., 2008; Knoth and Herrmannova, 2013). Thirdly, the method is also unable to suggest completely new pages for which a page does not exist yet. This means that if the collection is fairly small, the number of links that can be generated is also small.

A useful way to utilise semi-structured information has been presented by Mihalcea and Csomai (2007). They proposed a measure called *keyphraseness* expressing how likely it is for a term to act as a link. It is based on the number of documents where the term was already selected as a link ($count(D_{key})$) divided by the total number of documents in Wikipedia where the term occurs ($count(D_W)$).

$$P(keyword|W) \approx \frac{count(D_{key})}{count(D_W)} \tag{3.3}$$

This idea has been applied in a number of link discovery systems including the one we present in Section 7.2.

Granitzer et al. (2008) tried to use the anchor's context (defined, for example, as the sentence, section or document in which the anchor appears) together with information about page titles to create a disambiguation component that is not informed by the link graph. Their similarity measure is based on the ranking model of the Lucene search engine library[9].

3.2.2.3 Purely content-based link discovery

Purely content-based link discovery methods work only with plain text as input. They typically rely on some form of text-mining (NLP and information retrieval techniques) to discover new links, for example, keyword extraction,

[9] `http://lucene.apache.org/core/`

disambiguation and measuring semantic similarity. These methods are highly portable across domains and tasks. In a high proportion of cases, they are also the only available solution. On the other hand, they might be computationally expensive and provide generally lower performance than other methods.

The most typical application of these methods is probably in the context of document-to-document link discovery already described above. They have been used in many application domains to solve a wide variety of problems as we will show below. Surprisingly, in the context of evaluation conferences like INEX and NTCIR, these methods have not been widely tested. It has been argued by Knoth et al. (2011) that the development of purely content-based methods is not sufficiently encouraged by evaluation conferences. These methods are unlikely to score high and as the evaluation does not consider what information the method relies on, there is little incentive for researchers to submit them for evaluation.[10]

Zhang and Kamps (2009) submitted their content-based method to INEX:Link the Wiki Track. They employ classic IR techniques based on the Lucene implementation of the VSM model to combat the problem. Their algorithm works in two steps: 1) a fixed m number of documents are accepted in response to a query (the whole document text) as a set of candidate documents and 2) a fixed n number of iterative searches are carried out to find anchors and

[10]We will address this issue in Section 7.3.

generate links.[11] The authors conclude that even though their approach is not competitive with anchor-based (understand link-based) approaches, these methods are needed in a number of domains, such as to improve navigation in cultural heritage datasets.

3.2.2.4 Hybrid link discovery

Many of the latest link discovery methods rely on multiple types of input data. Sometimes, content-based methods are used as a baseline and the overall performance of the system is improved by replacing certain components of the system with semi-structured or link-based approaches, which can better exploit the specificities of the document collection.

Most of the systems submitted to the latest NTCIR:CrossLink evaluations can be seen as hybrid systems. This demonstrates that utilizing all types of information during the link discovery process leads to better performance than any of these approaches on its own. However, this also means that the systems are more complex and highly sensitive to changes of the task and environment.

[11] Although this method actually uses the title information in this step, it could be easily modified to depend on other information and thus we still consider this a content-based approach.

3.2.3 Link discovery methods according to use case and application context

In addition to the exploratory search (and discovery) use case, on which this thesis focuses, link discovery methods have been tested in the context of other related use cases. Some of the more prominent include:

- near-duplicate content detection (documents, phrases, quotations)
- plagiarism identification
- argument analysis
- citation analysis and bibliometrics (for example, recommending new citations for papers)

Link discovery methods can also be divided according to the domain in which they are applied. Some of the interesting domains include:

- digital libraries
- patent databases
- advertising
- profiling of individuals and recommendation

3.2.3.1 Near-duplicate content detection

Near-duplicate detection can be formulated as a task of finding documents or textual fragments that are almost identical. Near-duplicate detection can be seen as a special case of discovery, which is focused on finding content of the same or related origin. Traditionally, near duplicate detection is approached by using standard information retrieval techniques, such as using the vector space model with $tfidf$ weights and cosine similarity. Another approach is by using a technique called fingerprinting introduced by Rabin (1981) and originally used to detect unauthorised modification of documents.

Manku (2007) used Charikar's simhash (Charikar, 2002) to identify near-duplicate documents in a multi-billion page repository of Google. Simhash is a dimensionality reduction fingerprinting technique which has the property that fingerprints of near-duplicates differ in a small number of bits. This is, in principle, the opposite property of hash functions used to encrypt passwords where a small change in the input results in a significant change of the fingerprint.

Yang and Callan (2006) argue that detecting near-duplicates at a document level is insufficient. They have studied the problem of identifying specific classes of near-duplicates, which might not necessarily exhibit high similarity though they have the same origin. These include near-duplicate pairs which are created by adding or removing paragraphs, through modification or rewriting of some paragraphs, by repeating of a document as part of another, etc. One

of the main contributions of their work is that they try to use this additional knowledge to also identify the provenance of a near-duplicate.

3.2.3.2 Plagiarism detection

Plagiarism detection can be seen as a special case of near-duplicate detection. Consequently, they are closely related from a technical perspective. Plagiarism detection aims to detect documents that have significantly borrowed content from other documents. The task is complicated by the fact that these documents can often be of a very different length, the person plagiarising content might try to obfuscate the detection, for instance, by changing the vocabulary, the plagiarised copy might be in a different language, etc. An important forum for evaluation of plagiarism detection systems is PAN. In (Potthast et al., 2012), they have evaluated 18 plagiarism detection systems providing a shared framework for their comparative evaluation.

3.2.3.3 Argument analysis

A problem with organisation of (Web) resources, especially in specialised collections, is that the discovery and the modeling of arguments across resources is not sufficiently supported. There are many different reasons for modelling arguments spanning multiple resources. For instance, consider we want to analyse the public opinion on a recent political issue by connecting and collecting the opinions from discussions on the web, understand the scientific debate

on the issue of whether magnesium levels influence the risk of migraine, etc. Link discovery techniques can be applied with the aim to discover materials discussing a given topic, thus connecting the content on which argument modelling techniques can be applied. Link discovery can also assist in qualifying the relationships between individual materials using semantic labels, such as *contradiction*, *agreement* or *similarity*.

Taxonomies of such semantic labels (link types) have been developed by a number of researchers as discussed in Section 2.5. However, relatively few research papers dealt with the problem of detecting link types according to these taxonomies across documents and in an automated way. We will address this issue in Section 5.1.

3.2.3.4 Citation analysis and bibliometrics

The unprecedented growth of research papers worldwide and the availability of the vast majority of them online makes it possible to apply link discovery techniques to assist researchers in recommending relevant papers to read or even cite.

Caragea et al. (2013) developed a citation recommendation system based on Singular Value Decomposition (SVD). The idea is that the author provides an initial set of citation references. The system exploits the network of citations to recommend other citation references the author might have missed or should

be aware of. Strohman et al. (2007) on the other hand employs both textual and citation graph features to recommend relevant papers.

Interestingly, link discovery techniques can also be applied in the area of assessing research impact. Traditionally, impact metrics have been based primarily and purely on citations. In recognition of the trend of using citation counts for evaluating research excellence, Jiang, Zhuoren and Liu (2013) developed an algorithm for recovering missing citations in databases of scientific papers. The authors argue that, for example, 18.5% of articles in the ACM Digital Library are missing information on what they are citing, while 55.6% of articles are missing information about how often they are cited. The idea is to overcome the problem of poor connectivity of the graph for the purposes of ranking researchers. This is achieved by first exploiting the existing graph using the PageRank algorithm and then using co-authorship information to link researchers whose publications might be missing or not well represented in the citation graph.

Other researchers have argued that citations themselves are not a mark of quality and that if citations are used for evaluating research excellence, the citation type needs to be taken into account. A citation typing ontology CiTO has been, for example, developed by Shotton (2010). Other researchers, such as Teufel et al. (2006) or Bertin and Atanassova (2012), have worked on the problem of developing automatic citation typing tools based on information

extraction and machine learning techniques. A major problem for these tools is the lack of a database of research papers allowing access to both the full-text and citation information, so that these measures can be tested and applied at a global scale.

However, the mainstream work explored in this area is currently aiming at finding new metrics based on the ideas of *Webometrics* (Almind and Ingwersen, 1997) and *Altmetrics* (Priem et al., 2010). These measures, as well as the traditional citation measures, are based on the premise that the impact of a publication can be assessed outside of the publication space itself, that is by taking into account the scholarly debate in the form of citation counts, the usage statistics and the interactions on the social web as influenced by others. What is perhaps a bit surprising is that little work is following the idea that the full-text of the publication is the most important evidence for assessing its value, as started by the promising work on automatic citation typing. As discussed, a possible explanation to this might be the lack of a database of research papers allowing access to both the full-text and citation data at the time of the emergence and wide availability of usage and social data.

All these developments are interesting from the perspective of link discovery methods. For example, we can ask whether link discovery techniques can be developed to complement or even replace citation measures as the evidence of impact, as one of the major weaknesses of citation measures is that they can be

easily abused or gamed. The DiggiCORE project (Knoth, 2014) had the goal to apply link discovery techniques to create a large dataset of open access articles' full-texts accompanied with citation links as well as automatically constructed links between related papers. The intention has been to use this dataset as a basis for experiments on how semantic relatedness is correlated with citations. The work on *Semantometrics* reported in (Knoth and Herrmannova, 2014) goes in this direction.

3.2.3.5 Digital libraries

One of the most important areas where cross-document relations play a key role are digital libraries. Nowadays, the activities of researchers and students rely more and more on access to large online repositories using technologies and tools such as Google Scholar, CiteSeer or PubMed. While these systems are typically well suited for look-up search, i.e. finding relevant documents based on a keyword query, so far they have not sufficiently supported exploratory search. There is a range of use cases for link discovery methods in digital libraries. They include content recommendation at a document level, cross-document interlinking of content at a noun-phrase level, identification of near-duplicate content as an aid to a peer-review and identification of expertise based on authorship of similar content. Consequently, link discovery systems can be applied to improve the navigation capabilities in digital libraries.

3.2.3.6 Patent databases

Patent databases are another promising application domain for link-discovery methods. As patents are a legal protection excluding others from exploiting an invention, the strategy of patent applicants is sometimes to hide the patent application in hope of winning a future legal case, rather than announcing to others that a certain invention is protected. To achieve this effect, some patent applications use non-standard terminology or make it deliberately difficult to understand a patent. As a result of this, patent databases are often very difficult to search and navigate as these features are essentially obfuscating keyword search. Link-discovery methods can be used in patent databases to improve the exploration of the patent database, making it easier to check that an invention has not been patented.

3.2.3.7 Advertising

Online advertising is a fast-growing domain where link-discovery methods can be applied. The content recommendation use case at a document-to-document granularity is of a particular interest in this context. This includes, for example, recommending products to buy in online marketplaces. For example, Katukuri et al. (2013) developed a recommendation system for products on eBay. The system is based on an analysis showing that it is important to strike a balance between the similarity of the recommended products and

their quality. The authors show that their link discovery system improves the click-through rate, user engagement and revenue.

3.2.3.8 Recommendations based on user profiles

Another application domain for link-discovery methods, which partially overlaps with the advertising domain, is content recommendation based on user profiles. In this case, individuals, such as bloggers, researchers or newspaper readers, are represented by a set of texts they have produced or read. These texts are defining their user profile. The user profile is dynamic, i.e. changes according to the user's behaviour. Based on this profile, new content can be recommended to suit the specific needs of the users. The strength of this approach is particularly in the ability of a system to recommend newly created "fresh" material matching a user's interest without the need of the user to pro-actively search or subscribe to a certain topic.

3.3 Evaluation of link discovery systems

This section discusses common approaches to the evaluation of link discovery systems. While the evaluation of link discovery systems is based on traditional information retrieval measures, it is still important to understand how these measures are applied in the link discovery context, the advantages and disadvantages of the various measures and their potential modifications in the link

discovery context. Finally, it is also vital to be aware of the datasets on which systems can be tested and their desirable characteristics.

3.3.1 Traditional information retrieval evaluation measures and their use in link discovery

The traditional IR evaluation measures have been developed around the concept of *relevance* with respect to an information need. The set of all relevant answers a system can produce with respect to a specific input is referred to as the *gold standard* or *ground truth*.[12] All answers the system produces that are not in the ground truth are deemed as *nonrelevant*. The two most commonly used evaluation measures based on this concept are *precision* and *recall*. Using the definition from Manning et al. (2008), precision is defined as the fraction of the number of retrieved items that are relevant,

$$P = \frac{\text{number of relevant items retrieved}}{\text{number of retrieved items}} \tag{3.4}$$

while recall is defined as the fraction of the number of relevant documents that are retrieved.

$$R = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items}} \tag{3.5}$$

[12] At evaluation conferences (TREC, INEX and NTCIR), including link-discovery evaluations at INEX and NTCIR, an information need together with the ground truth set is called a *topic*.

A measure that provides a weighted mean of precision and recall is the *F-measure*, which is defined as

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}, \tag{3.6}$$

where $\alpha \in [0, 1]$. All these measures are set based, meaning that they allow us to assess the performance of the system given a set of answers the system produced. However, such approach is often not practical. Link discovery systems typically rank retrieved items, producing an ordered list where each item receives a score that should reflect its relevance to the information need. Consequently, the overall performance of the system typically changes depending on the number of items retrieved by the system (precision decreases and recall increases).

In this situation, we are often interested in assessing the performance of the system for the top k items that have been retrieved. We can therefore calculate precision and recall at k. Since in many situations, only the top few retrieved items matter, precision-at-k with k as low as 5 or 10 has been used in link discovery (Huang, Xu, Trotman and Geva, 2008; Huang et al., 2009; Tang, Geva, Trotman, Xu and Itakura, 2011; Tang et al., 2013; Trotman et al., 2009). To get a more complex view on the performance of the system, it is useful to create an *interpolated precision-recall curve*, which shows the maximum precision of the system at different recall levels. A widely used

evaluation measure is the N-point interpolated average precision. This allows one to produce a single interpolated precision-recall curve at N recall points (typically 0.0 to 1.0 with 0.1 intervals) by averaging precision at these points for multiple information needs. This approach has been used, for example, in the NTCIR CrossLink evaluations (see Chapter 7).

However, as there is the desire to simplify the assessment of a system's performance, *Mean Average Precision (MAP)* has widely been used at evaluation conferences to produce a single value by averaging the precision of a system at all recall levels. For a set Q of information needs, where for each $q_j \in Q$ there is a ground truth $\{d_1, \ldots, d_{m_j}\}$ and a ranked list R_{jk} of retrieval results, MAP is defined according to Manning et al. (2008) as:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1} m_j P(R_{jk}) \tag{3.7}$$

In the NTCIR CrossLink evaluation, the *Linked Mean Average Precision (LMAP)* was used to denote a measure similar to MAP where m_j is set to the number of identified items (250) instead of the size of the ground truth for all $q_j \in Q$.

Another measure used in link discovery as an alternative to MAP is *R-precision*. R-precision addresses the problem that the set of relevant documents might have a different size $|Rel|$ for each information need, thus even an ideal system might not be able to achieve precision score of 1.0 at k (when k is

higher than $|Rel|$) or recall at k (when k is smaller than Rel). R-precision is a precision at k where $k = |Rel|$ (typically averaged for all information needs in the collection). Consequently, an ideal system will always achieve R-precision 1.

3.3.2 Defining the ground truth in link discovery

Before one can start creating an evaluation framework, it is important to decide which answers of the system should be considered relevant and which not.[13] There are a number of options when deciding on the criteria for these binary judgements in link discovery. These options are based on the definition of a generated link and our understanding of relevance.

Using the definition from Section 3.1, we can understand a link as a pair $\langle s, t \rangle$ where s is the source and t is the target of the link. In evaluations, the source typically identifies a textual fragment, such as a noun phrase (sometimes called just *term*) or a whole document. This textual fragment, when used as a source of the link, is also sometimes referred to as *anchor*. The target might again represent any textual fragment. Let's say that we want to evaluate a link discovery system which identifies terms in one document and connects them with relevant paragraphs in other documents. A natural way how to define the set of relevant links would be to include only pairs generated by the

[13] There exist evaluation measures that can deal with non-binary (graded) relevance judgements. We will explain why they are needed later in this section and also in Section 7.3.

system where all of the following conditions are met:

- The boundaries of the anchor term are correctly identified.
- The boundaries of the target paragraph are correctly identified.
- The generated pair correctly contains the semantic relation of interest.

In practice, it might sometimes be difficult or impractical to enforce such a definition, because:

- *Strictness* — A potentially minor mistake in the detection of the boundaries of the link source or target results in the non-relevance of the whole link. Consequently, comparing systems might become difficult, as the evaluation does not take into account the differences in the types of errors systems make.
- *Effort needed to acquire the ground truth* — The work to create a ground truth for a data set with the precise boundaries of the link source and target can be tedious.
- *Subjectivity* — The decision on whether a link should or should not be established and also the definition of the boundaries of the link source and target can often be seen as fairly subjective.

One way to address these issues is to create multiple ground truths (for different criteria) and evaluate systems using them in parallel. This approach

has been taken in the INEX Link the Wiki Track and NTCIR CrossLink evaluations. For example, a successful answer of a noun phrase-to-document link discovery system is, with respect to the file-to-file (F2F) ground truth, any link connecting the correct documents, regardless of the correct identification of the noun phrase boundaries. However, the system is required to correctly identify both the link and the noun phrase boundaries to provide a correct answer with respect to the anchor-to-file (A2F) ground truth. This evaluation approach helps to better understand the types of errors systems make and acquire valuable information for improving them.

In some cases, a set of links is implicitly available in the collection. In such cases, it is often practical to re-use this information in the evaluation, as this data can be seen as an established ground truth. However, ground truth established in this way can often be incomplete, for example, due to subjectivity or personal preference of linking information. Consequently, one needs to be aware of these limitations when interpreting the results of such experiments. Care also needs to be taken in order not to over-fit systems with respect to incomplete and subjective ground truths.

The approach taken at INEX Link The Wiki Track and NTCIR CrossLink here is to use two types of assessments: a *manual* and an *automatic* assessment. The automatic assessment is done utilising the existing link structure in the collection. In this case, the ground truth is available prior to the experi-

ment. The manual assessment is established by pooling answers from different systems after the experiment and evaluating these answers by a set of (human) judges. Both evaluation approaches were at INEX and NTCIR run in parallel as each has its own advantages and disadvantages. The automatic approach is very useful in the stage of developing a system, as different parameters can be tested quickly for increased performance. It also allows to estimate the system's recall. On the other hand, it uses a potentially incomplete ground truth set and assessment at the automatic ground truth set might not be available at the desired granularity. The manual assessment can more easily work at the correct granularity, but is time consuming and typically done just once for systems comparison purposes. It also does not allow to estimate the overall recall, as we do not have assessments for all possible links in the dataset available.

While the limitations related to the incompleteness and subjectivity of the automatic ground truth have been acknowledged by the organisers of INEX Link The Wiki and NTCIR CrossLink Huang et al. (2009); Trotman et al. (2009), the impact this has on the results of experiments has not yet been quantified. A study carried out by Ellis et al. (1994) presents an experiment to measure the consistency of human subjects in inter-linking documents. Their study concluded that this consistency is generally low and at the same time variable for different data. This suggests that assessing the agreement on

a link discovery ground truth set is important, because the impact on the results might be substantial. One way to assess this is by measuring the inter-annotator agreement. In information retrieval, this is usually done using *Cohen's Kappa*, which is calculated as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{3.8}$$

where $Pr(a)$ is the relative observed frequency of agreement and $Pr(e)$ is the hypothetical probability of chance agreement. We will use this measure to assess the agreement between ground truths generated by different human subjects and to compare the performance of link discovery methods with the performance of humans in Chapter 6.

3.3.3 Datasets for link discovery evaluation

We will now have a look on the characteristics that a suitable dataset for evaluation of cross-document link discovery techniques should have:

- *Size* — A substantially large collection of textual documents.
- *Semantic relationship* — Links created as a reflection of a semantic relationship between the link source and target.
- *Correctness* — Links either created authoritatively (by an expert) or agreed by a community, so that there is only an insignificant proportion

of invalid links.

- *Completeness* — The number of correct relevant links, which can be discovered in the dataset, but are not already explicitly stated, is low.
- *Consistency* — The approach to the creation of links in different documents is consistent across the collection.

As we can see, the first two characteristics are derived from the definition of link discovery, while the last three characteristics follow directly from the 3C theorem presented in Section 2.3.3. Acquiring datasets with these characteristics is, in practice, very difficult. We will now list and document collections with a good potential to be used in evaluation of link discovery systems.

Wikipedia is a very popular collection for evaluation of link discovery systems. It is very large, freely available to download, written in multiple languages (making it applicable also for cross-lingual link discovery). It has a rich link structure which has been agreed by a large community of users. It was the collection of choice for INEX and NTCIR link discovery evaluations. Perhaps the main disadvantage of this dataset is that it contains only one type of explicit links. We can call them *conceptual* links, as they connect a concept to a document describing that concept. The encyclopedic nature of Wikipedia also dictates that conflicting statements should not be present on two different pages making the collection less useful for applications trying to detect those relationships. This dataset is widely used in this thesis in Chapters 4-7.

Research papers - Research papers potentially constitute an excellent corpora for evaluation of link discovery system. Research articles form a network connected by citation links created by their authors. The advantage with research articles is that they provide an opportunity to test many different types of relations, including also discourse types of relationships. Although there are millions of research papers available on the Internet, text-mining of the majority of these articles has not been legal in the UK until June 2014 when a copyright exception for text-mining for non-commercial and research purposes recommended by the Hargreaves (*Implementing the Hargreaves review*, 2014) came into effect. As of today, machine access to many of these articles is still restricted (Knoth, 2013; Knoth, Anastasiou and Pearce, 2014). To the best of our knowledge, there has been no sufficiently large dataset that would contain both citation links and full-texts of research papers that could be exploited for experimenting with link discovery methods. However, it is widely believed that the solution here will be brought by the Open Access movement, which promotes both access and re-use of research papers. We will discuss in detail our contribution to this movement and the development of an open dataset for these purposes in Chapter 8.

Bible - A potentially useful collection for link discovery. The bible is probably the most read book of all time (*The Bible Tops the List of the Most Read Books in the World*, 2014) and has been translated to the vast majority

of languages. Bible concordances, i.e. manually curated indexes to improve the access to bible passages, created as early as in the 13th century can be seen as predecessors of modern indexing systems and hypertext. For example, the Scofield Reference Bible first published in 1909 has a cross-referencing system which connects related verses and events across different books and chapters of the bible. For example, the Hypertext Bible project makes all the texts available for download to developers (*Bible Data Files*, 2014). Although the bible has many characteristics of a suitable collection for evaluation of link discovery systems, surprisingly, we are not aware of any such use in this area.

Newspaper collections - Newspaper articles constitute, similarly to research papers, excellent data for evaluation of link discovery systems. Online newspaper articles often contain links to related articles. While there are databases providing access to digitised (typically early 20th century) newspapers, unfortunately, to the best of our knowledge, there is no dataset that would provide access to modern newspapers (due to copyright restrictions) in a form that could be text-mined and would contain cross-references between articles.

Overall, there is no ideal collection that would provide a widely accepted standard for the evaluation of link discovery systems. However, up to some extent and despite certain limitations, the Wikipedia collection plays this role at evaluation conferences. While the other mentioned collections have the

potential to be used in link discovery evaluations, their application has been problematic due to a variety of reasons, including restrictions on access and the difficulty to extract explicit links to be used as the ground truth.

3.3.4 Alternative approaches to evaluation of link discovery

One of the shortcomings of the traditional evaluation measures described in Section 3.3.1 is the definition of the concept of relevance (according to which a document either is or is not relevant). In reality, it is difficult to provide binary relevance judgements as the outputs provided by a system can be more or less relevant, i.e. certain links can be very relevant, others might be moderately relevant and the rest slightly relevant or not relevant at all. If relevance judgements are binary, this can lead to situations in which, for example, a good strategy for a system to perform well might be to rank results according to the confidence of the system rather than the relevance of the results to the information need. Järvelin and Kekäläinen (2002) were one of the first to propose new evaluation metrics based on the concept of *graded relevance*, which is based on the assumption that highly relevant documents are more valuable than marginally relevant documents. A thorough comparison of binary and graded relevance measures can be found in (Sakai, 2009).

Other approaches to evaluation of link discovery methods can be based on

performing a user-centered evaluation rather than evaluation using standard measures. For example, Blustein (1999) evaluated his system by measuring how much time users save when looking for a specific information with and without automatically generated links. In practice, these studies are extremely time consuming to perform and consequently do not allow for optimisation of the system's parameters based on the evaluation results.

3.4 Conclusions

This chapter provided a formal definition of the link discovery task and reviewed the state of the art. We have learned that there is a wide variety of link discovery methods according to the granularity of the link source and target and the applied use case. We have also discussed a range of approaches to link discovery with respect to their input data and their implications on performance and applicability across collections. As it is problematic to comment on the performance of link discovery approaches that rely on different input data and datasets and operate at a different granularity, we stress the importance of comparative evaluations, such as those provided by INEX: Link The Wiki Track and NTCIR CrossLink. Since the evaluation of link discovery methods is key to the progress in this area, we also reviewed existing evaluation approaches, identified suitable evaluation datasets and highlighted some of the current challenges in the comparative evaluation of link discovery systems.

The work presented in this chapter also helped us to identify further research gaps, opportunities and challenges, which influenced the formulation of the research sub-questions (RQ 1 - RQ 4) and goals presented in Section 2.7.2. They include the potential to study the relationship between semantic similarity and the behaviour of humans in linking content, the opportunity to use multilingual corpora to test the disparity of links created by different communities, the need for the development of text-based link discovery methods and the difficulty of applying link discovery techniques on datasets where exploratory search experiences are desirable. Finally, reviewing this work had an impact on the design of link discovery methods presented in the following chapters.

Chapter 4

To Link or Not To Link: The Study of Human Linking Behaviour in Wikipedia and its Relation to Semantic Similarity

In sciences, researchers often study various phenomena trying to describe them using mathematical models. Similarly, to develop a link discovery system, i.e. a method that can detect links between textual fragments, we should first study the human behaviour of linking textual content. The aim of the work presented in this chapter is to better understand the properties of content that is linked by people using hypertext and its implications for the link discovery task. The chapter addresses the following research question:

RQ 1: *Is it possible to identify properties that would indicate which pairs*

of textual resources are more likely to be linked by people?

Text retrieval methods are typically designed to find documents relevant to a query based on some criterion, such as Okapi BM25 or cosine similarity (Manning et al., 2008). Similar criteria have also been used to identify documents relevant to a given reference document, thus, in principle, carrying out document-to-document link discovery. A number of these approaches use measures of semantic similarity. However, the correspondence of these measures to the way people link content has not been sufficiently investigated (see Chapters 1 and 3). As our contribution to this topic, we study the predictive potential of semantic similarity for automatic link discovery. We do this by investigating this correspondence on a large text corpus and by designing a method based on the outcomes of this analysis.

As part of our work, we also take a closer look at the impact of the length of documents on predictive power of semantic similarity. This is motivated by the fact that when a collection contains long documents, better retrieval performance is often achieved by breaking each document into subparts or passages and comparing these rather than the whole documents to a query (Manning et al., 2008). A suitable granularity of the breakdown is dependent on a number of circumstances, such as the type of the document collection and the information need. Consequently, we have decided to investigate link discovery at the level of documents and paragraphs and have developed a

fairly simple two-step paragraph-to-paragraph link discovery method, which draws on the knowledge we acquired from the initial analysis. It consists of the following steps:

1. Given a collection of documents, our goal is to identify candidate pairs of documents between which a link may be induced.

2. Given each candidate pair of documents, our task is to identify pairs of passages, such that the topics in the passages are related in both documents.

The rest of the chapter is organised as follows: Section 4.1 discusses the data selected for our experiment and Section 4.2 describes how the data were processed in order to perform our investigation. In Section 4.3, the analysis in which we compared the results produced by semantic similarity measures with respect to the way people link content is presented. Section 4.4 then draws on this analysis and introduces the link discovery method which is finally evaluated in Section 4.5. We provide a summary of the original contribution of this chapter in Section 4.6.

4.1 Data selection

The following properties were required for the document collection to be selected for the experiments. First, in order to be able to measure the correlation

between the way people link content and the results produced by semantic similarity measures, it was necessary to select a document collection which can be considered as relatively well interlinked. Second, it was important for us to work with a collection containing a diverse set of topics so that the outcomes could be generalised. Third, we required the collection to contain articles of varied length. Nevertheless, we were mostly interested in longer documents, which create conditions for the testing of passage retrieval methods. We decided to use the Wikipedia collection, because it satisfies all these requirements (see Section 3.3.3).

The English version of Wikipedia consists of more than four million pages spread across five hundred thousands categories. As it would be unnecessarily expensive for our calculation to work with the whole encyclopedia, a smaller, but still a sufficiently large subset of Wikipedia, which satisfies our requirements of topic diversity and document length, was selected. Our document collection was generated from articles in categories containing the phrase "United Kingdom." This includes categories, such as United Kingdom, Geography of United Kingdom or History of the United Kingdom. There are about 3,000 such categories and 57,000 distinct articles associated to them. As longer articles provide better test conditions for passage retrieval methods, we selected the 5,000 longest articles out of these 57,000. This corresponds to a set where each article has the length of at least 1,280 words.

4.2 Data preprocessing

Before discussing the analysis performed on the document collection, let us briefly describe how the documents were processed and the semantic similarity calculated.

First, the N articles/documents $D = \{d_1, d_2, \ldots, d_N\}$ in our collection were preprocessed to extract plain text by removing the Wiki markup. The documents were then tokenised and a dictionary of terms $T = \{t_1, t_2, \ldots, t_M\}$ was created. Assuming that the order of words can be neglected (the bag-of-words assumption) the document collection can be represented using a $N \times M$ term-document matrix. In this way, each document is modelled as a vector corresponding to a particular row of the matrix. As it is inefficient to represent such a sparse vector in memory (most of the values are zeros), only the non-zero values were stored. *Term frequency—inverse document frequency (tfidf)* weighting was used to calculate the values of the matrix. Term frequency tf_{t_i,d_j} is a normalised frequency of term t_i in document d_j:

$$tf_{t_i,d_j} = \frac{f(t_i, d_j)}{\sum_k f(t_k, d_j)} \tag{4.1}$$

Inverse document frequency idf_{t_i} measures the general importance of term t_i in the collection of documents D by counting the number of documents

which contain term t_i:

$$idf_{t_i} = \log \frac{|D|}{|d_j : t_i \in d_j|} \tag{4.2}$$

$$tfidf_{t_i,d_j} = tf_{t_i,d_j}.idf_{t_i} \tag{4.3}$$

Similarity is then defined as the function $sim(\overrightarrow{x}, \overrightarrow{y})$ of the document vectors $\overrightarrow{x}$ and $\overrightarrow{y}$. There exists a number of similarity measures used for the calculation of similarity between two vectors (Manning and Schuetze, 1999), such as *cosine*, *overlap*, *dice* or *Jaccard* measures. Some studies employ algorithms for the reduction of dimensions of the vectors prior to the calculation of similarity to improve the results. These approaches may involve techniques, such as lexical chaining (Green, 1999), Latent Semantic Indexing (Deerwester, Dumais, Furnas, Landauer and Harshman, 1990), random indexing (Widdows and Ferraro, 2008) and Latent Dirichlet Allocation (Blei et al., 2003). In this work we intentionally adopted perhaps the most standard similarity measure — cosine similarity calculated on the *tfidf* vectors and no dimensionality reduction technique was used. The formula is provided for completeness:

$$sim_{cosine}(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x}.\overrightarrow{y}}{|x|.|y|} \tag{4.4}$$

Cosine similarity with *tfidf* vectors has been previously used in automatic link discovery systems producing state-of-the-art results when compared to

other similarity measures (Chen et al., 2004). This allows us to report on the effectiveness of the most widely used measure with respect to the way the task is completed by people. While more advanced techniques might be in some cases better predictors for link discovery, we did not experiment with them as we preferred to focus on the investigation of the correlation between the most widely used measure and manually created links. Such a study has to our knowledge never been done before, but it is necessary for the justification of automatic link discovery methods.

4.3 Semantic similarity as a predictor for link discovery

The document collection described in Section 4.1 has been analysed as follows. First, pair-wise similarities using the formulas described in Section 4.2 were calculated. Cosine similarity is a symmetric function and, therefore, the calculation of all inter-document similarities in the dataset of $5,000$ documents requires the evaluation of $\frac{5{,}000^2-5{,}000}{2} = 12,497,500$ combinations. Figure 4.1 shows the distribution of the document pairs (on a log_{10} scale) with respect to their similarity value. The frequency follows a power law distribution. In our case, 99% of the pairs have similarity lower than 0.1. It is possible to see a small spike with a peak in the region with similarity of around 0.9. We believe

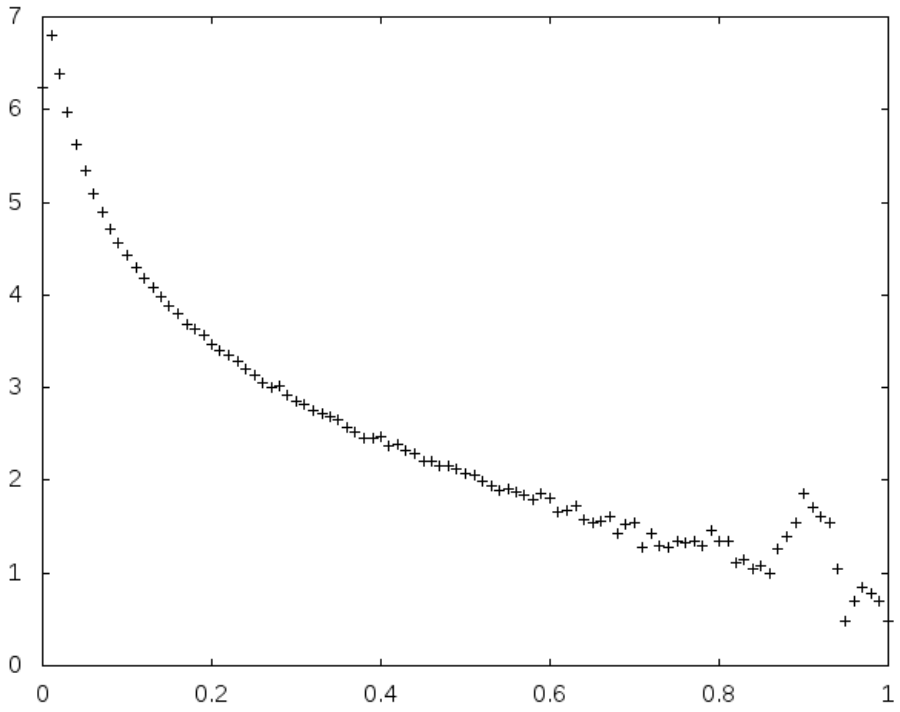


Figure 4.1: The histogram shows the number of document pairs on a $\log_{10}$ scale (y-axis) with respect to their cosine similarity (x-axis).

that this might be due to the Wikipedia collection containing pages that follow similar discourse patterns, such as *Transport in London* and *Transport in Manchester*. As we will be normalising with respect to the number of samples in each similarity region, we believe the shown distribution does not affect the further reported results.

To compare the semantic similarity measures with the links created by Wikipedia authors, all inter-document intra-collection links, i.e. links created by users of Wikipedia commencing from and pointing to a document within our collection, were extracted. These links represent the connections as seen by the

users regardless of their direction. Each of these links can be associated with a similarity value calculated in the previous step. Documents with similarity lower than 0.1 were ignored. Out of the 120,602 document pairs with inter-document similarity higher than 0.1, 17,657 pairs were also connected by a user-created link.

For the evaluation, interval with cosine similarity $[0.1, 1]$ was divided evenly into 100 buckets and all 120,602 document pairs (samples) were assigned to the buckets according to their similarity values. From the distribution shown in Figure 4.1, buckets corresponding to higher similarity values contain fewer document pairs than buckets corresponding to smaller similarity values. Therefore, for each bucket, the number of user created links within the bucket was normalised by the number of document pairs in the bucket. This number is the likelihood of the document pair being linked and will be called *linked-pair likelihood*. The relation between semantic similarity and linked-pair likelihood is shown in Figure 4.2.

As reported in Chapters 2 and 3, semantic similarity has been previously used as a predictor for the automatic discovery of links. The typical scenario was that the similarity between pairs of documents was calculated and the links between the most similar documents were generated (Wilkinson and Smeaton, 1999). If this approach was correct, we would expect the curve shown in Figure 4.2 to be monotonically increasing. However, the relation shown in Figure 4.2

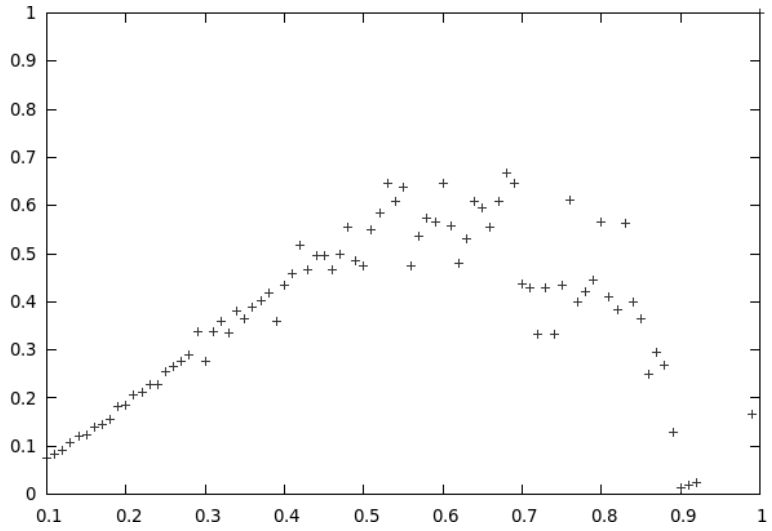


Figure 4.2: The linked-pair likelihood (y-axis) with respect to the cosine similarity (x-axis).

is in accordance with our expectations only up to the point 0.55. For higher values of inter-document similarity the linked-pair likelihood does not rise or it even decreases. We have observed a similar trend also using another document collection and report on these results in Section 8.4.4.

Spearman's rank correlation and Pearson correlation were applied to estimate the correlation coefficients and to test the statistical significance of our observation. This was performed in two intervals: $[0, 0.55]$ and $[0.55, 1]$. Very strong positive correlations 0.986 and 0.987 have been received in the first interval for the Spearman's and Pearson coefficients respectively. Negative correlations -0.640 and -0.509 have been acquired for the second interval again for the Spearman's and Pearson coefficients respectively. All the measured correlations are significant for p-value well beyond $p < 0.001$.

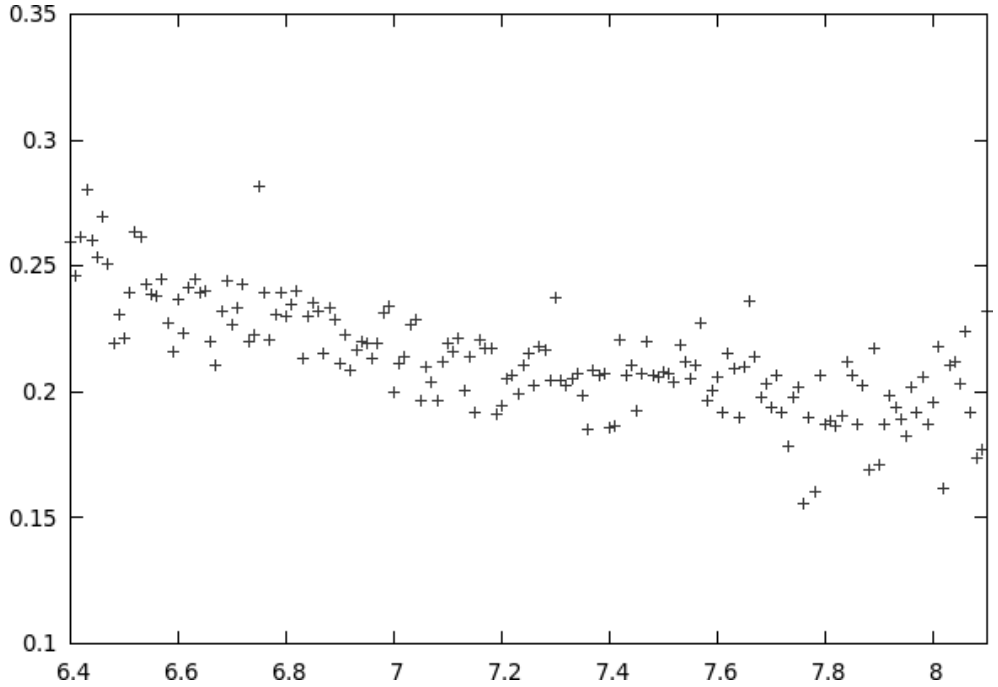


Figure 4.3: The average cosine similarity (y-axis) of document pairs of various length (x-axis) between which there exists a link. The x-axis is calculated as a $\log_{10}(l_1.l_2)$

The results indicate that high similarity value is not necessarily a good predictor for automatic link discovery. A possible explanation for this phenomenon is that people create links between related documents that provide new information and therefore do not link nearly identical content. However, as content can be in general linked for various purposes, more research is needed to investigate if document pairs at different similarity levels also exhibit different qualitative properties. More specifically, can the value of semantic similarity be used as a predictor for relationship typing? We investigate this question in Chapter 5.

An important property of semantic similarity as a measure for automatic

discovery of links is the robustness with respect to the length of documents. As mentioned in Section 4.2, cosine similarity is by definition normalised by the product of the documents length. Ideally the cosine similarity should be independent of the documents length. To verify this in our dataset, we have taken pairs of documents between which Wikipedia users assigned links and divided them into buckets with respect to the function $\log_{10}(l_1.l_2)$, where l_1 and l_2 are the lengths of the two documents in the document pair and the logarithm is used for scaling. The value of each bucket was calculated as an average similarity of the bucket members. The results are shown in Figure 4.3. The graph shows that the average similarity value is slightly decreasing with respect to the length of the articles. Values -0.484 and -0.231 were obtained for Spearman's and Pearson correlation coefficients respectively. Both correlations are statistically significant for $p < 0.001$. A much stronger correlation was measured for Spearman's than for Pearson which can be explained by the fact that Spearman's correlation is calculated based on ranks rather than real values and is thus less sensitive to outliers.

Our experience from repeating the same experiment on another Wikipedia subset generated from categories containing the word Geography tells us that the decrease is even more noticeable when short and long articles are combined. The decrease in average similarity suggests that if cosine similarity is used for the automatic discovery of links then document pairs with higher value of

$l_1.l_2$ have a higher linked-pair likelihood than pairs with a smaller value of this quantity. In other words, links created between documents with small $l_1.l_2$ typically exhibit a larger value of semantic similarity than links created between documents with high value of $l_1.l_2$. Although the decrease may seem relatively small, we believe that this knowledge may be used for improving automatic link discovery methods by adaptively modifying the thresholds with respect to the $l_1.l_2$ length.

4.4 Link discovery method

In this section we introduce the method for the automatic discovery of links. The method can be divided into two parts (1) Identification of candidate link pairs (i.e. the discovery of document-to-document links) (2) Recognition of passages sharing a topic between the two documents (i.e. the discovery of passage-to-passage links).

4.4.1 Document-to-document links

The algorithm for link discovery at the granularity of a document is motivated by the findings reported in Section 4.3.

The algorithm takes as the input a set of document vectors and two constants – the minimum and maximum similarity thresholds – and iterates over all pairs of document vectors. It outputs all document vector pairs, such that

Algorithm 1 Generate document links

Input. Set of document vectors D, min. sim. α, max. sim. $\beta \in [0,1]$, $C = \emptyset$
Output. Set C of candidate links of form $\langle d_i, d_j, sim \rangle \in C$ where d_i and d_j are documents and $sim \in [0,1]$ is their similarity

1: **for all** $\{\langle d_i, d_j \rangle | i, j \in \aleph_0 \wedge i < j < |D|\}$ **do**
2: $sim_{d_i,d_j} :=$ SIMILARITY$((d_i, d_j))$
3: **if** $sim_{d_i,d_j} > \alpha \wedge sim_{d_i,d_j} < \beta$ **then** $C := C \cup \langle d_i, d_j, sim_{d_i,d_j} \rangle$
4: **end if**
5: **end for**.

their similarity is higher than α and smaller than β. For well chosen β, the algorithm does not generate links between nearly duplicate pairs. If we liked to rank the discovered links according to the confidence of the system, we would suggest to assign each pair a value using the following function.

$$rank_{d_i,d_j} = |sim_{d_i,d_j} - (\alpha + \frac{\beta - \alpha}{2})| \tag{4.5}$$

The ranking function makes use of the fact that the system is most confident in the middle of the similarity region defined by constants α and β, under the assumption that suitable values for these constants are used. The higher the rank of a document pair, the better the system's confidence.

4.4.2 Passage-to-passage links

Due to a high number of combinations, it is typically infeasible even for relatively small collections to generate passage-to-passage links across documents directly. However, the complexity of this task is substantially reduced when

passage-to-passage links are discovered in a two-step process.

Algorithm 2 Generate passage links

Input. Sets P_i, P_j of paragraph document vectors for each pair in C, min. sim. γ, max. sim. $\delta \in [0,1]$ such that $\alpha < \gamma \wedge \beta < \delta, L = \emptyset$

Output. Set L of passage links of form $\langle p_{k_i}, p_{l_j}, sim \rangle \in L$ where p_{k_i} and p_{l_j} are paragraphs in documents d_i, d_j and $sim \in [0,1]$ is their similarity

1: **for all** $\{\langle p_{k_i}, p_{l_j} \rangle | p_{k_i} \in P_i, p_{l_j} \in P_j\}$ **do**
2: $\quad sim_{p_{k_i}, p_{l_j}} :=$ SIMILARITY$((p_{k_i}, p_{l_j}))$
3: $\quad$ **if** $sim_{p_{k_i}, p_{l_j}} > \gamma \wedge sim_{p_{k_i}, p_{l_j}} < \delta$ **then** $L := L \cup \langle p_{k_i}, p_{l_j}, sim_{p_{k_i}, p_{l_j}} \rangle$
4: $\quad$ **end if**
5: **end for**.

As Section 4.3 suggests, the results of Algorithm 1 may be improved by adaptive changing of the thresholds α and β based on the length of the document vectors. More precisely, in the case of cosine similarity, this is the quantity $lr = l_1.l_2$. The value α should be higher (β lower) for pairs with low lr than for pairs with high lr and vice versa. Although the relative quantification of this ratio is left for future work, we believe that we can exploit these findings for the discovery of passage-to-passage links.

More specifically, we know that the length of passages (paragraphs in our case) is lower than the length of the whole documents. Hence, the similarity of a linked passage-to-passage pair should be on average higher than the similarity of a linked document-to-document pair, as revealed by the results of our analysis. This knowledge is used within Algorithm 2 to set the parameters γ and δ. The algorithm shows, how passage-to-passage links are calculated

for a single document pair previously identified by Algorithm 1. Applying the two-step process allows the identification of document pairs, which are likely to contain strongly linked passages, at lower similarity levels and to recognise the related passages at higher similarity levels while still avoiding duplicate content.

4.5 Results

The experimental evaluation of the methods presented in Section 4.4 is divided into two parts: (1) the evaluation of document-to-document links (Algorithm 1) and (2) the evaluation of passage-to-passage links (Algorithm 2).

4.5.1 Evaluation of document-to-document links

As identified in Section 4.3 (and shown in Figure 4.2), the highest linked-pair likelihood does not occur at high similarity values, but rather somewhere between similarity 0.5 and 0.7. According to Figure 4.2, the linked-pair likelihood in this similarity region ranges from 60% to 70%. This value is in our view relatively high and we think that it can be explained by the fact that Wikipedia articles are under constant scrutiny by users who eventually discover most of the useful connections. However, how many document pairs that could be linked in this similarity region have been missed by the users? That is, up to what extent can our method help in the discovery of possible connections?

Suppose that our task would be to find document pairs about linking of which the system is most certain. In that case we would set the thresholds α and β somewhere around these values depending on how many links we would like to obtain. In our evaluation, we have extracted pairs of documents from the region between $\alpha = 0.65$ and $\beta = 0.70$ regardless of whether there originally was a link assigned by Wikipedia users. An evaluation tool which allowed an evaluation participant[1] to display the pair of Wiki documents next to each other and to decide whether there should or should not be a link between the documents was then developed. We did not inform the participant about the existence or non-existence of links between the pages. More specifically, the evaluation participant was asked to decide yes (link generated correctly) if and only if they found it beneficial for a reader of the first or the second article to link them together regardless of the link direction. The evaluation participant was asked to decide no (link generated incorrectly) if and only if they thought that navigating the user from or to the other document does not provide additional value. For example, in cases where the relatedness of the documents is based on their lexical rather than their semantic similarity.

The study revealed that 91% of the generated links were judged by the evaluation participant as correct and 9% as incorrect. Table 4.1 shows the results of the experiment with respect to the links originally assigned by the users of Wikipedia. It is interesting to notice that in 3% of the cases the

[1]A colleague in the department who agreed to this evaluation.

		Wikipedia link	
		yes	no
Subject's decision	yes	61%	30%
	no	3%	6%

Table 4.1: Document-to-document links from the $[0.65, 0.7]$ similarity region. The subject's decision in comparison to the Wikipedia links.

subject decided not to link the articles even though they were in fact linked on Wikipedia. Overall, the algorithm discovered in 30% of the cases a useful connection which was missing in Wikipedia. This is in line with the findings of Huang, Trotman and Geva (2008) who claims that the validity of existing links in Wikipedia is sometimes questionable and useful links may be missing.

An interesting situation in the evaluation occurred when the subject discovered a pair of articles with titles *Battle of Jutland* and *Night Action at the Battle of Jutland*. The Wikipedia page indicated that it is an orphan (a page without any links pointing to it) and asked users of Wikipedia to link it to other Wikipedia articles. Our method would suggest the first article as a good choice.

4.5.2 Evaluation of passage-to-passage linking

The previous section provided evidence that the document-to-document linking algorithm is capable of achieving high performance when parameters α, β are well selected. However, Section 4.3 indicated that it is more difficult to discover links across long document pairs. Thereby, we have also evaluated

		Wikipedia link	
		yes	no
Subject's decision at page level	yes	16%	10%
	no	18%	56%

Table 4.2: Document-to-document candidate links discovery from the $[0.2, 0.21]$ similarity region and document pairs with high lr ($lr \in [7.8 - 8]$).

		System's decision	
		yes	no
Subject's decision	yes (correct)	14%	46%
	no (incorrect)	24%	16%

Table 4.3: Passage-to-passage links discovery for very long documents. Passages extracted from the $[0.4, 0.8]$ similarity region.

the paragraph-to-paragraph linking on document pairs with quite low value of similarity $[0.2, 0.21]$. According to Figure 4.2, this region has only 15% linked-pair likelihood.

Clearly, our goal was not to evaluate the approach in the best possible environment, but rather to check whether the method is able to discover valuable passage-to-passage links from very long articles with low similarity. Articles with this value of similarity would be typically ranked very poorly by link discovery methods working at the document level.

Table 4.2 shows the results after the first step of the approach, described in Section 4.4, with respect to the links assigned by Wikipedia users. As in the previous experiment, the evaluation participant was given pairs of documents and decided whether they should or should not be linked. Parameters α and

β were set to 0.2, 0.21 respectively. Table 4.2 indicates that that the accuracy ($16\% + 10\% = 26\%$) is at this similarity region much lower than the one reported in Table 4.1, which is exactly in line with our expectations. It should be noticed that 34% of the document pairs were linked by Wikipedia users, even though only 15% would be predicted by linked-pair likelihood shown in Figure 4.2. This confirms that long document pairs exhibit a higher probability of being linked in the same similarity region than shorter document pairs.

If our approach for paragraph-to-paragraph link discovery (Algorithm 2) is correct, we will be able to process the document paragraphs and detect possible paragraph-to-paragraph links. The selection of the parameters γ and δ influences the willingness of the system to generate links. For this experiment, we set the parameters γ, δ to 0.4, 0.8 respectively. The evaluation participant was asked to decide: (1) if the connection discovered by the link discovery method at the granularity of passages was useful (when the system generated a link) and (2) whether the decision not to generate a link is correct (when the system did not generate a link). The results of this evaluation are reported in Table 4.3. It can be seen that the system made in 60% ($14\% + 46\%$) of the cases the correct decision. Most mistakes were made by generating links that were not sufficiently related (24%). This might be improved by using a higher value of γ (lower value of δ).

4.6 Summary of contribution

The aim of this chapter was to explore whether it is possible to identify certain properties that are indicative of texts being more likely to be linked by people. We showed that there is a statistically significant correlation between the probability of text pairs being linked (linked-pair likelihood) and the value of semantic similarity, measured as cosine similarity using $tfidf$ term-document vectors. We explained that this correlation (see Figure 4.1) is interesting as it shows that the linked-pair likelihood is directly proportional to semantic similarity only up to a certain threshold. With a higher semantic similarity, linked-pair likelihood starts decreasing as evidenced by a negative correlation.

This information is valuable for the development of text-based link discovery systems that are aimed at improving navigation, such as content recommendation systems. Our findings suggest that recommendations should not be ranked descendingly according to the value of semantic similarity, as it would be the case in traditional look-up search. We have used this knowledge for the development of a novel two-step algorithm for link discovery at the granularity of documents and paragraphs, which ranks document pairs according to their expected linked-pair likelihood instead of just semantic similarity (Section 4.4). This finding has also been applied in the CORE system we present in Chapter 8.

Our experiments also confirm that the length of documents is an important factor influencing the performance of using semantic similarity as a predictor for link discovery. They show that links between long documents with short but related passages are more difficult to discover in this way. This justifies the development of methods working at a finer granularity than documents. Such methods are currently not widely used in practice. In the mainstream search engines, they are not even used in the look-up scenario, yet there is a potential they could significantly improve the speed of access to information, in particular in the case of long documents. The results also suggest that it might be possible to improve the performance of link discovery methods in the future by considering the length of the texts in the ranking phase.

Overall, our main original contribution is that we provided a new insight into the use of semantic similarity as a predictor for automatic link discovery by performing an investigation in the way people link content. This motivated us in the development of a novel purely content-based approach for automatic discovery of links at the granularity of both documents and paragraphs, which does not expect semantic similarity and linked-pair likelihood to be directly proportional.

Chapter 5

The Meaning of a Link: Using Semantic Similarity as a Criterion for Cross-Document Link Typing

In the previous chapter, we explored how people link textual content. In particular, we focused on analysing the relationship between the value of semantic similarity of two texts and the probability people connect these texts by a link. Our results suggest linked-pair likelihood can be used as a predictor to decide if two texts should or should not be connected by a link. This link is of an unspecified type, exactly as a hypertext link. However, being able to recognise and assign semantic types (see Section 2.5) to (hypertext) links would be useful for improving the performance of link discovery methods and their adaption

to specific use cases. It could even lead to the emergence of new innovative applications. While the current hypertext specification does not support link typing (hypertext links do not have a semantic type), this does not constitute a barrier for investigating this problem nor potentially applying it in practice (as workarounds are available).

There has been a significant research effort in the area of modelling cross-document relationships (see Section 2.5). They include semantic relations at the discourse level ranging from mere similarity of topics presented in two documents to the assertion that one document elaborates/contradicts the ideas described in another one. Enriching document collections by cross-document relationships provides the means for better organising fragmented information. It can help improve the browsing, the navigation and the discovery of important information. However, the current cross-document relationship modelling approaches rely on human annotators and therefore do not scale in large constantly growing document collections. So far, little work has addressed these limitations.

The work presented in this chapter builds on the results described in Chapter 4, exploring whether the value of semantic similarity is indicative of the link (relationship) type. We investigate the different types of cross-document relationships, explore which of these types might be possible to detect automatically using semantic similarity as a criterion and discuss the implications

and the application areas of automatic link typing methods. The chapter addresses the following research question:

RQ 2: *Can some of the properties identified by answering RQ 1 be used to suggest the semantic type of a link?*

The rest of the chapter is organised as follows. In Section 5.1, we quickly reflect on the work in cross-document document link typing, making a case for automatic link typing methods. Building on Chapter 4, we present an experiment investigating the relationship between the value of semantic similarity and a selected set of link types (Section 5.2). We summarise the original contribution of this chapter in Section 5.3.

5.1 Towards automatic assignment of link types

The idea of typed cross-document links has been first introduced by Trigg (1983) as part of his hypertext taxonomy of link types (see Section 2.5). Since that time, a number of new, typically domain specific, taxonomies have been developed (Buckingham Shum et al., 2000; Mancini, 2005; Radev et al., 2008). All this work has been motivated by the aim to enable the reuse of knowledge, which can be created by analysing, comparing and contrasting information from multiple documents.

One way to look at the work of authoring typed cross-document links is to see it as the process of semantic modelling of cross-document discourse/argument.

This argument, represented by a web of typed relationships, is typically not explicitly stated in any single document. It is the end-product of interpreting information from multiple documents.

In a technical sense, the creation of an individual typed cross-document link can be seen as the generation of Type 3 metadata as defined in Section 2.3. In Trigg's work (Trigg, 1983), it is expected that these metadata are manually authored. Today, various social annotation tools for metadata generation, such as for image tagging, have become very popular on the Web. Applications using them are based on the assumption that a large number of users is capable of providing the necessary metadata in sufficient time and quality. In the light of this opportunity, Buckingham Shum and Ferguson (2010) expected that applying a collective intelligence approach to model argumentation links across open educational resources, as implemented in the Cohere system (Buckingham Shum and De Liddo, 2010), will result in a user-generated web of meaningfully connected annotations, which can be visualized, filtered and searched for patterns in ways that are impossible at present.

However, as we know from Section 2.3, the success of these approaches is largely pre-determined by the metadata types to be generated. In the case of Type 3 metadata, the number of possible connections increases quadratically with respect to the number of resources. This creates a problem that is particularly significant in large, quickly growing document collections with many

contributing authors writing about different issues or in different languages. As a result, people are unable to keep track of all the potentially relevant information and connections (see Section 2.3.2). Increasing the number of people is not a solution to the problem, meaning that such an approach can only be successful in document collections of a very limited size. This creates the need for automatic link discovery and typing tools to assist in the process.

The problem of automatic link typing has been addressed by Allan (1996), who created a taxonomy of link types that he believes can be recognised using text-mining techniques. His methods are based on the idea of dividing documents into smaller textual fragments and calculating similarity between these fragments across documents, generating links when the similarity is higher than a threshold. The hypothesis is that the pattern of these generated links, characterised by the mutual position of links as demonstrated in Figure 5.1, could be used to suggest the link type. The shortcoming of this study is that it lacks any quantitative evaluation of the approach.

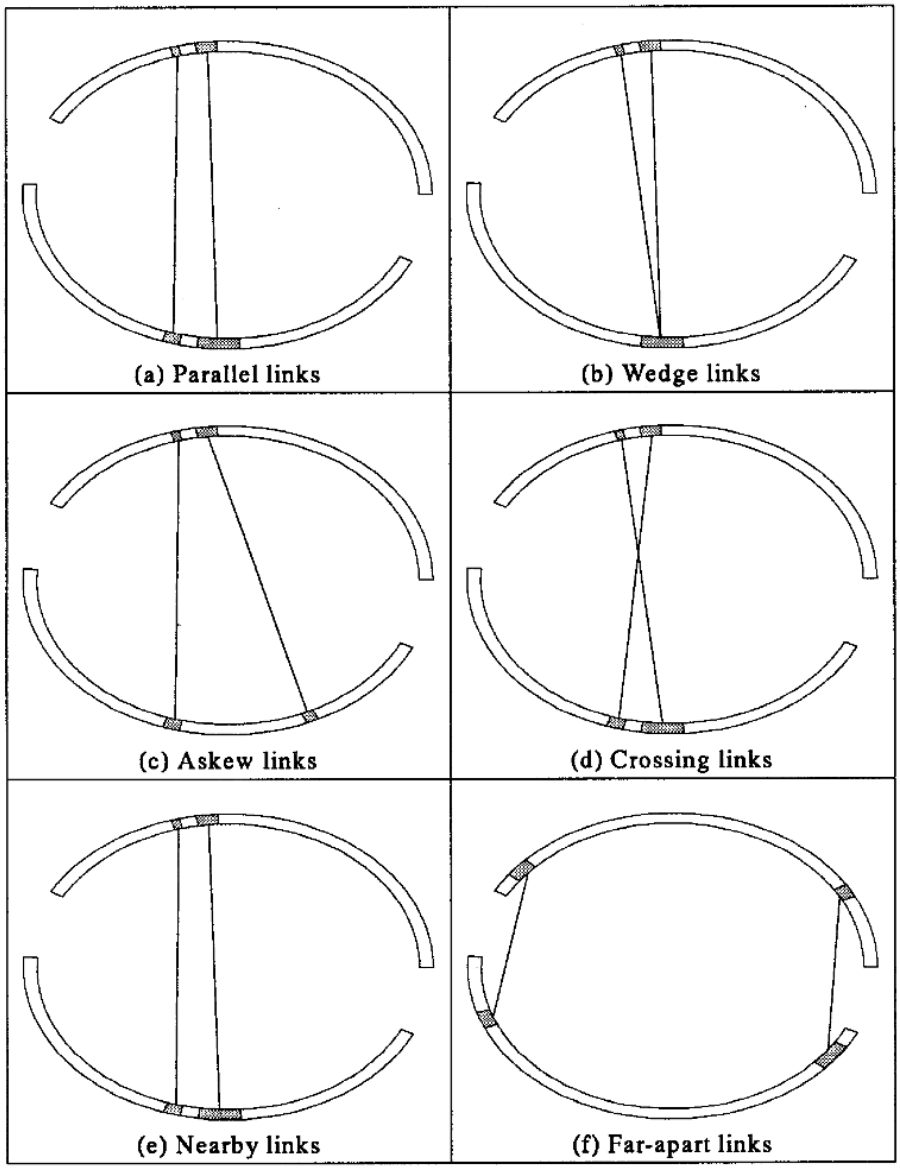


Figure 5.1: The hypothesis reported in (Allan, 1996) is based on the assumption that the mutual position of automatically generated links between short textual fragments of two documents is indicative of their semantic relationship type. The image, showing the types of mutual positions, is taken from (Allan, 1996).

5.2 Studying the potential of semantic similarity in link typing

In the previous chapter, we have studied the relation between links authored by people and semantic similarity. In this chapter, we are extending this work by investigating the qualitative properties of links using the same Wikipedia dataset as described in Section 4.1. The dataset has a number of the usual advantages of Wikipedia, such as the rich link structure agreed by a large number of people. Links in Wikipedia may represent different semantic relationships. However, only a limited set of discourse relations are present in Wikipedia at the article level. As a consequence, we do not investigate relations, such as disagreement or contradiction, which in Wikipedia should only appear in an explicit way at an intra-document level (contradiction across Wikipedia articles is not desirable and the role of the reviewers is to remove it).

5.2.1 Semantic similarity and linked-pair likelihood

A central concept of our study is the quantity called *linked-pair likelihood* introduced in Chapter 4, which is the probability that a pair of documents is connected by a manually created link, calculated as $lpr = \frac{|\text{links}|}{|\text{document pairs}|}$. Figure 4.2 shows *lpr* calculated for groups of document pairs at different intervals of semantic similarity. While it can be observed that linked-pair likelihood

strongly correlates with the value of semantic similarity, the direction of the correlation in the right part of the graph is quite unexpected, provoking the following questions:

(1) Why is the curve in Figure 4.2 not monotonically increasing after a certain similarity is reached? Why is it not true in the whole range of values that the more semantically similar two resource are, the more likely they are to be linked?

(2) As content can be linked for various reasons, are there any qualitative differences between linked documents with different value of semantic similarity?

A possible explanation for question (1) is that people create links between related documents that provide new information and therefore do not link nearly identical content. Regarding question (2), we hypothesise that the value of semantic similarity might be used in link type identification, i.e. the reasons for linking articles with different values of semantic similarity are also different.

5.2.2 Relations of interest and their representation

In our experiment, we have decided to use four discourse link types building on the classification provided by (Allan, 1996). We hypothesize that the value

of semantic similarity might be a useful discriminative factor for each of these link types. The sampled document pairs were classified to the following types: *tangent*, *similarity/equivalence*, *expansion*, *aggregate*. Examples of these link types are depicted in Table 5.1. The description of these link types is as follows:

Tangent links represent according to (Allan, 1996) links which relate topics in an unusual manner, for example, a link from a document about "Clouds" to one about Georgia O'Keeffe (who painted a mural entitled *Clouds*). In our work tangent links are associated to document pairs that are related in a useful, but relatively marginal way, typically there is a single piece of information that justifies the relationship of the documents.

Expansion link type is attached to a link which starts at a discussion of a topic and has as its destination a more detailed discussion of the same topic.

Similarity/equivalence links represent related and strongly-related discussions of the same topic.

Aggregate links are those which group together several related documents. According to Allan (1996), aggregate links may in fact have several destinations, allowing the destination documents to be treated as a whole when desirable. In our work, only pairs of documents are considered and

Title 1	Title 2	Link type	Description
Jack McConnell	Scottish Qualifications Authority	tangent	The first article mentions that the Scottish Labour politician Jack McConnell appointed a new board for the Scottish Qualifications Authority (SQA) and introduced significant changes to the way the agency worked.
Social Democratic Party (UK)	David Owen	expansion	David Owen was was one of the founders of the British Social Democratic Party (SDP) and led the SDP from 1983 to 1987 and the re-formed SDP from 1988 to 1990. The first article mentions David Owen a number of times.
Senior Railcard	Family and Friends Railcard	similarity/ equivalence	Both articles describe the history of railcards introduced by British Rail. Articles clearly describe two semantically related concepts.
Statutory Instruments of the UK, 1996	Statutory Instruments of the UK, 1996 (3001-4000)	aggregate	The first article contains the other as its part.

Table 5.1: Example link types

thus aggregate links are assigned to document pairs when the first article contains significant parts of the second article. The aggregate relationship is not an inverse of expansion as it does not connect a more detailed explanation/discussion of the topic addressed in the first document, but it rather refers to the reuse of certain pieces of text across documents.

The only discourse link types from Allan's taxonomy that we did not use for classification are *comparison* and *contrast* links. Contrast and comparison is

in a Wiki typically handled either explicitly in the text, e.g., *"The invasion of Iraq was particularly controversial, as it attracted widespread public opposition and 139 of Blair's MPs opposed it."* or it is part of the elaboration, revision and refinement process of the article. This obviously reduces the number of discourse relationships we can identify to those mentioned above. We also assume that two contrasting text segments would often be represented by similar term-document vectors and therefore the value of semantic similarity would not provide sufficient information to distinguish them.

5.2.3 Link typing results

To answer the questions specified in Section 5.2.1, we have carried out a study that investigates the characteristics of link pairs at different similarity levels. The interval $[0.1, 1]$ of semantic similarity, depicted in Figure 4.2, has been divided into 9 intervals of even width. As a case study, 10 article pairs from each interval[1] between which a link was created by Wikipedia users were randomly sampled and they were assessed by a human investigator and classified. This process resulted in obtaining 95 sample document pairs. An evaluation environment was created to allow the investigator to see the articles next to each other and to easily compare them. The investigator was asked to inspect both articles, to assign exactly one of the four relationships of interest and to

[1]Only 5 article pairs were sampled from the interval $[0.9, 1.0]$ due to lack of data in this region.

provide a brief justification for the decision. The document pairs were presented to the investigator in a random order and the investigator was during the evaluation not aware of the calculated value of semantic similarity associated with the article pairs. The evaluation and classification of one pair took from 5 to 20 minutes. The whole manual evaluation took about 19 hours.

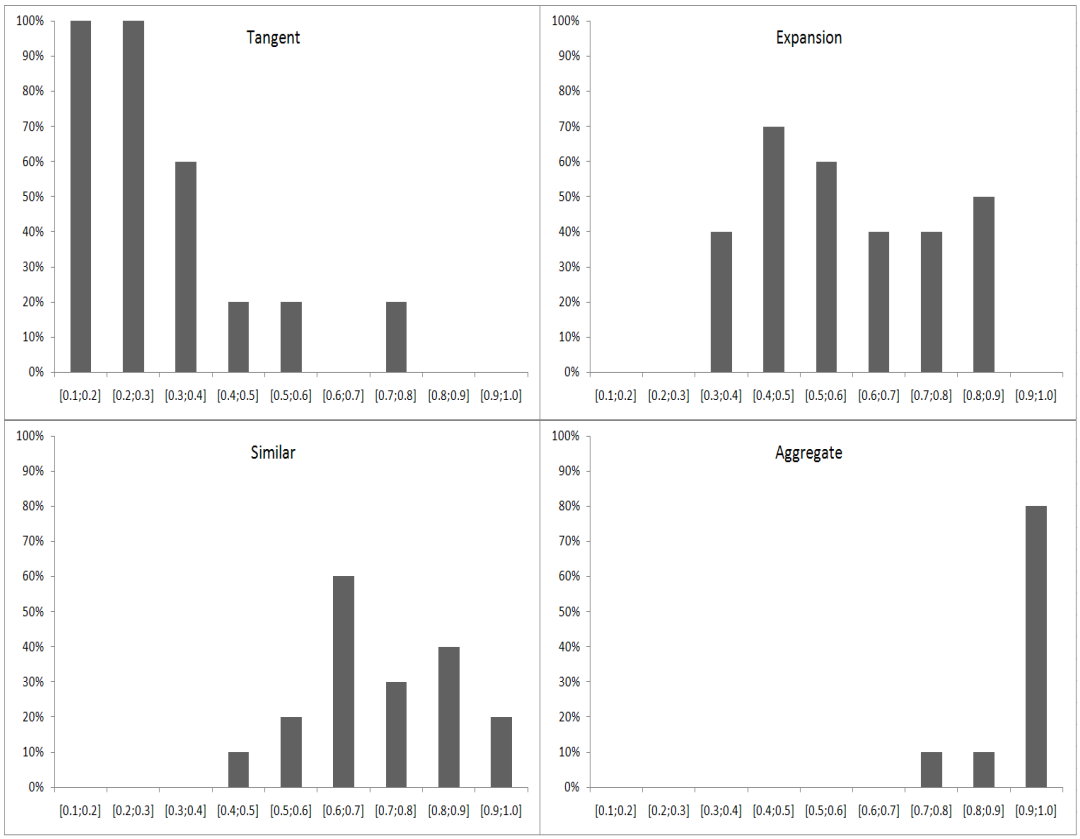


Figure 5.2: The frequency of different link types with respect to semantic similarity of document pairs

Overall, 37% of article pairs were classified as *tangent*, 36% as *expansion*, 20% as *similar* and 7% as *aggregate*, corresponding to the prior probability of

an article pair expressing a given relationship on this dataset. The results of the evaluation are presented in Figure 5.2. The figure shows the frequency of different link types in all the 9 selected intervals. This data could be used to assess the posterior probability of a link type given a specific value of semantic similarity. For example, it directly follows from Figure 5.2 that a document pair with similarity 0.3 has a 60% chance of being a tangent link and a 40% chance of being an expansion link.

We have found that in the lower levels of semantic similarity $[0.1, 0.3]$ most of the links were classified under the tangent link type. At higher levels of similarity the proportion of the tangent link types decreases. Only very few links were classified as tangent when the similarity of the document pair was high.

Expansion links start to appear at similarity higher than 0.3. At the similarity level of $[0.3, 0.4]$ the proportion of the expansion links is roughly the same as the proportion of tangent links. The highest proportion of expansion links is present in the semantic similarity interval of $[0.4, 0.6]$ where the value of similarity seems to be quite a distinctive factor from the similarity link types. At higher similarity values, the proportion of expansion links drops and similar link types appear.

Most of the similar/equivalence links types are present in the interval $[0.6, 0.9]$. The proportion of this link type is in this region approximately

40%. It seems that it is hard to distinguish them in this interval from the expansion links solely based on the similarity value. When semantic similarity reaches the value of 0.9, it is possible to see aggregate link types that are characteristic by a high similarity value.

Overall, this confirms that the value of semantic similarity is a useful factor characterising, to a certain extent, the type of semantic relationship. This provides an answer to the second question reported in Section 5.2.1. We have also observed from this experiment and Figure 4.2 that people link most often document pairs of the expansion and tangent types, even though the tangent type is in absolute numbers the most frequent link type. People link less likely document pairs providing similar, equivalent or even duplicate content.

The value of semantic similarity is just one criterion useful for the detection of link types, but has not been used in link typing previously. We expect that robust link typing systems should, however, combine multiple strategies to detect link types. We are aware that the value of semantic similarity, as presented in this example, is unable to make distinctions about certain link types, such as the *contrast* link type, nor it can be used to determine the direction of the link, for link types, such as *prerequisite*. Other text characteristics, perhaps combined with external knowledge, need to be tested for this purpose.

5.3 Summary of contribution

We have argued that automatic link typing systems are needed in order to provide scalable solutions that would facilitate the reuse of knowledge created by analysing, comparing and contrasting information from multiple documents. We argued that typed cross-document links in large document collections cannot be simply produced by the "social web," as is the case for some other metadata types. However, we believe it would be possible to confirm or reject automatically identified link types using crowd-sourcing approaches.

The main contribution of this chapter is that we showed, in our experimental study, that the value of semantic similarity is a useful indicator that can help to identify link types. We have used Wikipedia as a source of textual documents in our experiment, which allowed us us to simplify the problem by considering only a limited set of cross-document relations. We assume that it is possible that there exist more indicators, complementary to the value of semantic similarity, the combination of which could enable the development of more sophisticated link typing methods capable of recognising additional link types from Allan's taxonomy.

Chapter 6

Crossing the Language Barriers: New Methods for Document to Document Cross-Lingual Link Discovery (CLLD) and Evaluation

In the previous two chapters, we explored the link discovery problem on a dataset written in one language. In the monolingual setting, automatic link discovery methods are needed to ensure that data in large textual collections can be inter-linked (and kept inter-linked) efficiently. This improves the accessibility of information in these document collections and can facilitate the discovery of (hidden) knowledge. Studying and applying link discovery methods in the multilingual context is exciting for the following reasons:

- Assisting in the process of the discovery of semantically related infor-

mation written in any language emphasises the difference between discovery and look-up (search). While staying aware of semantically related information in a specific and fairly small domain might be possible (though tedious), the multilingual environment provides an extra (knowledge/skills) barrier further limiting the human ability of relying on look-up search when dealing with exploratory tasks.

- Assuming that information is linked because of the relatedness of semantics, one might expect different (language) communities to create similar link structures. The multilingual environment provides an opportunity to study the similarity/disparity of this process on multilingual collections with comparative texts.

In this chapter, we will explore how to automatically generate links between related documents written in different languages. Capitalising on the opportunities of the multilingual environment, we will also explore the similarity/disparity in the way speakers of different languages link content.

The chapter addresses the following research questions:

RQ 3: *How can we detect links between textual resources written in different languages?*

RQ 4: *How shall we interpret the performance achieved by link discovery methods and how does the technology compare to the ability of humans to carry out the same task?*

In order to address RQ 3, we present new methods for Cross-Lingual Link Discovery (CLLD) and provide their comparative evaluation on the Wikipedia corpus. Although our experiments are conducted on this dataset, we believe the results are applicable more widely as we use Wikipedia articles only as a set of general documents,[1] abstracting from the Wikipedia encyclopedic nature. In order to address RQ 4, we explore the agreement of human annotators linking articles in different language versions of Wikipedia. To investigate up to what extent can CLLD methods match the performance of humans, we also compare this agreement with the results achieved by our automatic CLLD methods.

The content of the chapter is organised as follows. We describe the CLLD methods in Section 6.1. Section 6.2 provides further details about the dataset used in our experiments and Section 6.3 introduces the issues in evaluating CLLD methods. We then describe the experimental setup, provide methods evaluation and present our agreement investigation in Section 6.4. Finally, we summarise the contribution of this chapter in Section 6.5.

6.1 The CLLD methods

Our methods are based on Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007). ESA is a method that calculates semantic related-

[1]This is not the case in Chapter 7 where we exploit the characteristic features of the Wikipedia collection.

ness of two texts by mapping their term-document vectors to a high dimensional space. This is typically, but not necessarily, the space of Wikipedia concepts. Each Wikipedia concept corresponds to a specific Wikipedia page and each dimension of the resulting ESA vector corresponds to one such concept/page. The value of the vector in each dimension expresses the similarity of the source text to that Wikipedia concept. The method then calculates the similarity between these high-dimensional vectors, instead of using standard term-document vectors. The projection of term-document vectors is done using a "semantic interpreter" which should be created from a large background corpus. The semantic intepreter takes as an input an N-dimensional vector, where N corresponds to the size of the vocabulary used in the input texts, and produces an M-dimensional vector, where M corresponds to the number of different documents in the background corpus. Each dimension in the output vector represents the similarity of the input document to one document in the background corpus. The ESA projection and similarity calculation process is schematically shown in Figure 6.1. The method has received much attention in the recent years and it has also been extended to a multilingual version called Cross-Lingual Explicit Semantic Analysis (CL-ESA) (Sorg and Cimiano, 2008*a*). To the best of our knowledge, this method has not yet been applied in the context of automatic link discovery systems.

As described in Section 3.2.2, current approaches to link discovery can be,

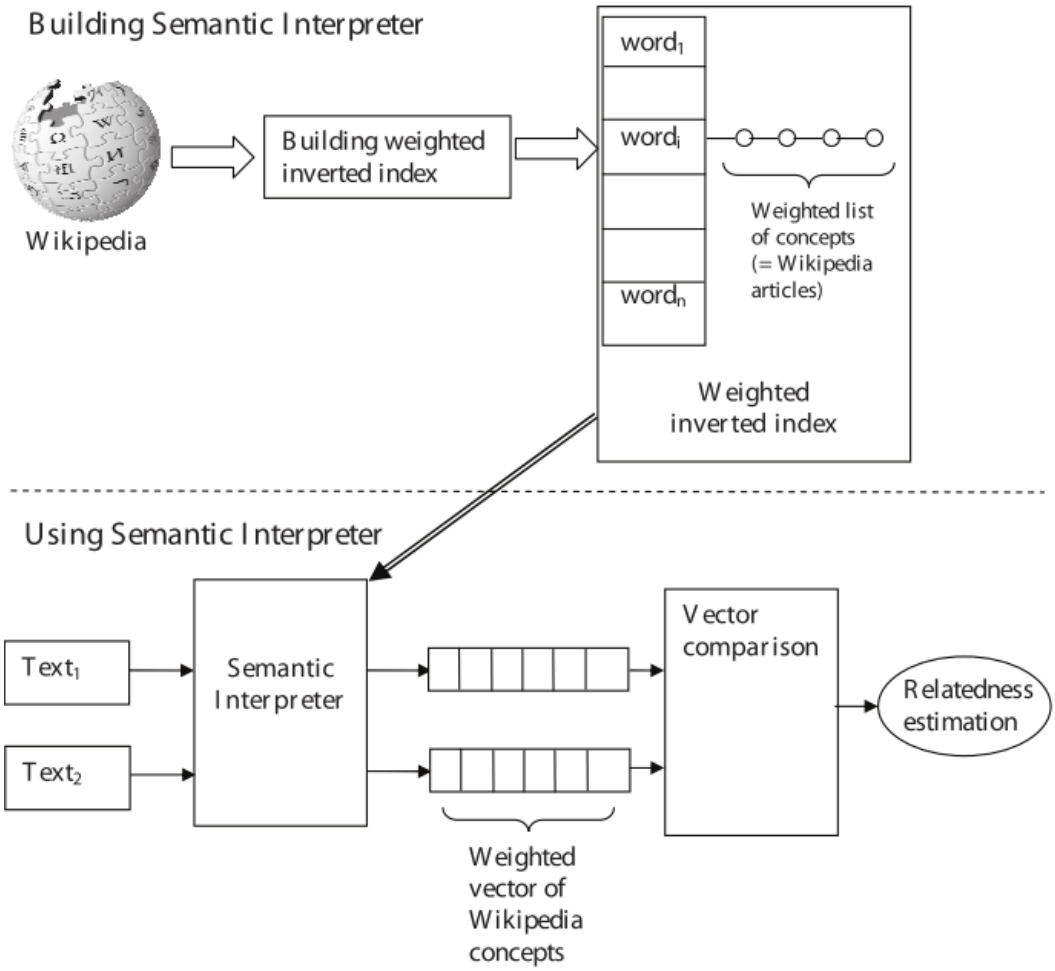


Figure 6.1: The projection of term-document vectors to a high-dimensional space using a "semantic interpreter" and the subsequent calculation of semantic similarity according to ESA. The image is taken from Gabrilovich and Markovitch (2007).

based on the use of input data, divided into link-based, semi-structured, purely text based and hybrid. In this chapter, we present four CLLD methods (three purely text-based and one combining the link-based and content-based approach). Measuring semantic similarity using ESA has been previously shown to produce better results than calculating it directly on document vectors using cosine and other similarity measures. Gabrilovich and Markovitch (2007) also found ESA to outperform the results that can be obtained by measuring similarity on vectors produced by Latent Semantic Analysis (LSA). Therefore, the cross-lingual extension of ESA seems a plausible choice for our work.

The overall CLLD process is demonstrated in Figure 6.2. Each method takes as an input a new "orphan" document (i.e. a document that is not linked to other documents) written in the source language and automatically generates a ranked list of documents written in the target language (the suitable link targets from the source document). The task involves two steps: the *cross-language* step and the *link generation* step. We have experimented with four different CLLD methods: *CL-ESA2Links*, *CL-ESADirect*, *CL-ESA2ESA* and *CL-ESA2Similar* that will be described later on. The names of the methods are derived from the approach applied in the first and the second step. The methods have different characteristics and might be useful in different scenarios.

In the **first step**, an ESA vector is calculated for each document in the

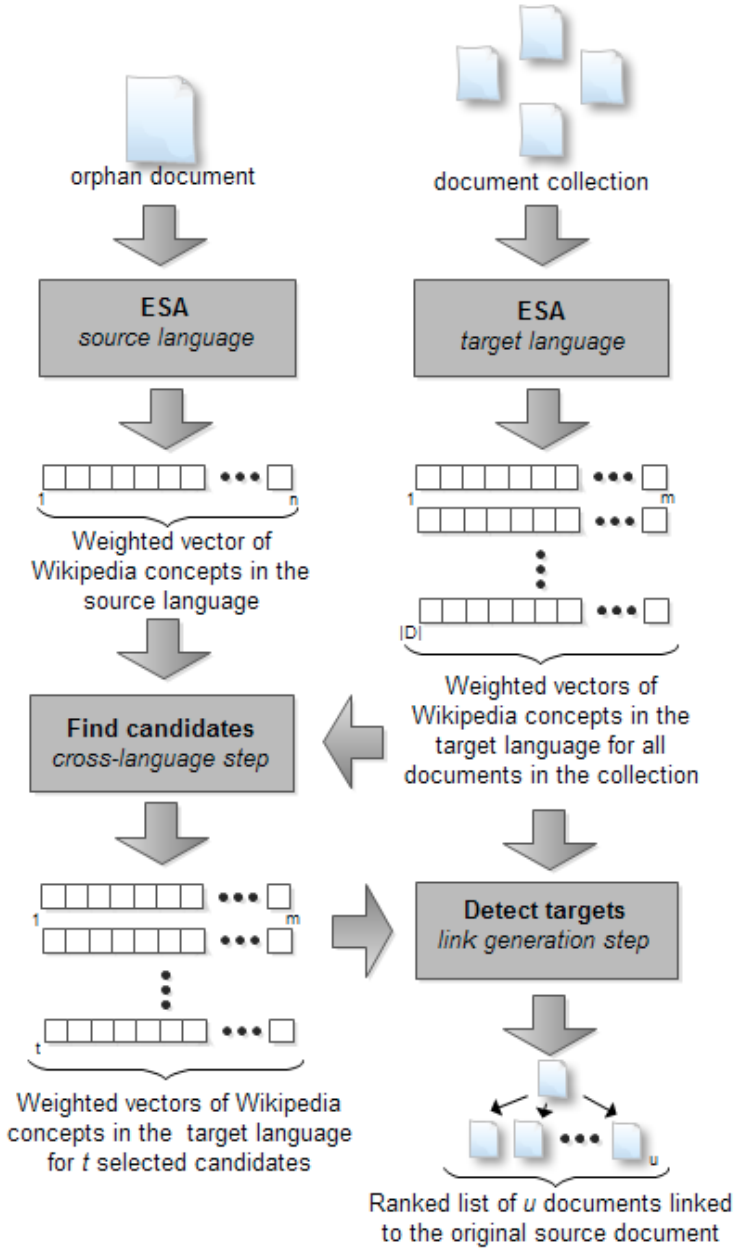


Figure 6.2: Cross-language link discovery process

document collection. This results in obtaining a weighted vector of Wikipedia concepts for each document in the target language. The cardinality of the vector is given by the number of concepts (pages) in the target language version of Wikipedia (i.e. it is about 3.8 million for English, 764,000 for Spanish, etc.). A similar procedure is applied on the orphan document, however, the source language version of ESA is used. The resulting ESA vector is then compared to the ESA vectors that represent documents in the target language collection (CL-ESA approach). A set of candidate vectors representing documents in the target language is acquired as an output of the cross-language step, see Section 6.1.1.

In the **second step**, the candidate vectors are taken as a seed and are used to discover documents that are suitable link targets. The four different approaches used in this step distinguish the above mentioned methods, see Section 6.1.2.

6.1.1 The cross-language step

The main rationale for the cross-language step is to find t suitable candidates in the target language that can later be exploited to identify link targets. Semantically similar target language documents to the source language document are considered by our methods as suitable candidates. To identify such documents, the ESA vector of the source document is compared to the ESA

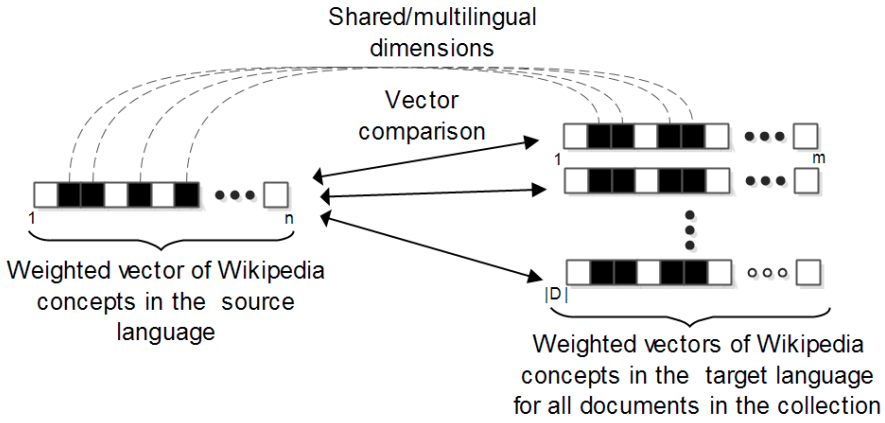


Figure 6.3: CLLD candidates

vectors of documents in the target document collection.

Each dimension in an ESA vector expresses the similarity of a document to the given language version of a Wikipedia concept/article. Therefore, the cardinality of the source document vector is different from the cardinality of the vectors representing the documents in the target language collection (Figure 6.3). In order to calculate the similarity of two vectors, we map the dimensions that correspond to the same Wikipedia concepts in different language versions (Figure 6.3). In most cases, if a Wikipedia concept is mapped to another language version, there is a one-to-one correspondence between the articles in those two languages. However, there are cases when one page in the source language is mapped to more than one page in the target language and vice versa.[2] For the purpose of similarity calculation, we use 100 dimensions with

[2]These multiple mappings appear quite rarely, e.g. in 5,889 cases out of 550,134 for Spanish to English and for 2,528 cases out of 163,715 for Czech to English.

the highest weight that are mappable from the source to the target language. The number of candidates to be extracted is controlled by parameter t. We have experimentally found that its selection has a significant impact on the performance of our methods.

6.1.2 The link generation step

In the link generation step, the candidate documents are taken and used to produce a ranked list of targets for the original source document. The following approaches, schematically illustrated in Figure 6.4, are taken by our four methods:

CL-ESA2Links — This method requires access to the link structure in the target collection. More precisely, the method takes the original orphan document in the source language and tries to link it to an already interlinked target language collection. After applying CL-ESA in the first step, existing links are extracted from the candidate documents to determine possible link targets. Reoccurring targets are treated in the same way as if they appeared just once. The link targets are then ranked according to their similarity to the source document, i.e. documents that are more similar are ranked higher. This list is then used as a collection of link targets.

CL-ESADirect — This method applies CL-ESA on the source document

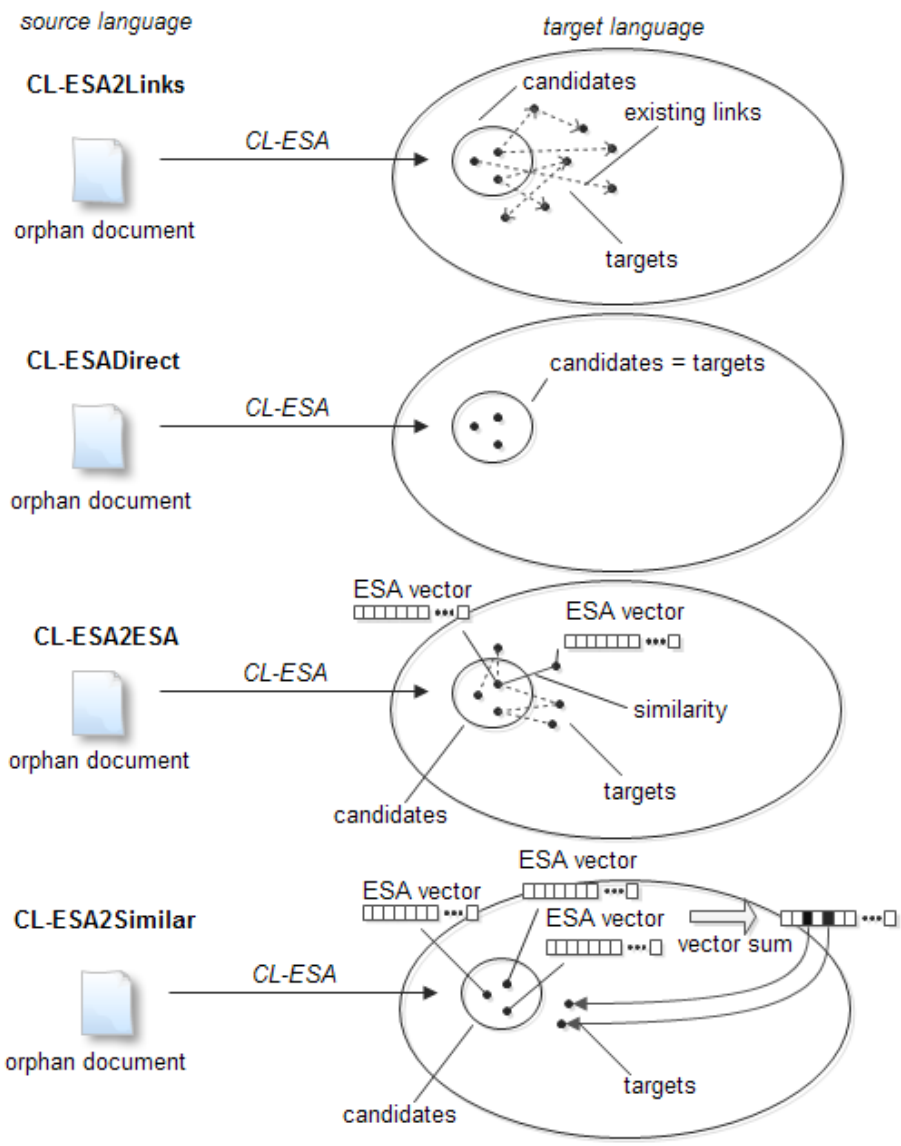


Figure 6.4: Schematic illustration of the four approaches used by the CLLD methods.

and takes the list of candidates directly as link targets.

CL-ESA2ESA — In this method, the application of CL-ESA is followed by another application of monolingual ESA, which measures the semantic similarity of the candidates with all documents in the document collection, to identify link targets.

CL-ESA2Similar — Instead of generating the ranked list of link targets using monolingual ESA, as in the previous method, which is computationally expensive, we calculate a vector sum from the candidate list of ESA document vectors. We then select strong Wiki concepts (as those representing dimensions with a high value of this sum) as the set of targets. This is equivalent to calculating cosine similarity using *tfidf* vectors. Though much quicker, the main disadvantage is that if we wanted to use this method on another set than Wikipedia, ESA would have to be used with a different background collection.

All of the methods have different properties. CL-ESA2Links requires the knowledge of the link graph in the target document collection. ESADirect and CL-ESA2ESA are two methods that are universal, i.e. can be easily applied in any document collection. The difference between them is that the former one requires significantly less document vector comparisons than the later method. CL-ESA2Similar works almost as fast as CL-ESADirect, but

it has the disadvantage that ESA has to be used with the specific document collection as a background.

6.2 The underlying data

Wikipedia has been selected as a suitable corpus for the methods evaluation (see Section 3.3.3 for details).

In our study, we have experimented with the English, Spanish and Czech language versions of Wikipedia. We consider the cases of linking from Spanish to English and from Czech to English, i.e. from a less resourced language to the more resourced one. We believe that this is the more interesting direction for CLLD methods as the target language version is more likely to contain relevant information not available in the source language. The language selection has been motivated by the aim to test the methods in two very different environments. The Spanish version is relatively well resourced containing 764,095 pages (about four times fewer than English), the Czech language is much less resourced containing 196,494 pages (about four times fewer than Spanish).[3]

[3]The mentioned figures refer to the size of the data dumps used in the experiment, i.e. the latest version of Wikipedia data dumps at the time of carrying out the experiment.

6.3 Evaluation methodology

From our definition of link discovery (see Section 3.1), we can see that the reason for linking two pieces of information is made at the level of semantics, i.e. the annotator has to understand the concepts/ideas described in two texts to decide if they should be connected by a link. We claim that this process should be language independent. Thus, an article about London will be related to an article about the United Kingdom regardless of the language the articles are written in.

Therefore, we can specify our task in the following way: Given a document in the source language, find documents in the target language that are suitable link targets for the source document, i.e. there is a semantic relationship between the source document and the linked target documents.

Based on this specification, the ground truth for a topic document d is the set of documents that can be considered (semantically) suitable link targets. Though this set is typically unknown to us, we can, in our experiment, approximate it by taking the existing Wikipedia links as a ground truth. Because the Wikipedia link structure has been agreed by a large number of contributing authors, one can assume it should have a relatively consistent link structure in comparison to content that would be linked just by a single person. To establish the ground truth for the original source document, we can extract

all links originating in the source document and pointing to other documents. Since the process of linking information is performed at the semantic level, and is thus language independent, we can enrich our ground truth with link graphs from different language versions of Wikipedia. This approach can be considered as the current state-of-the-art in the evaluation of link discovery systems and has been used by NTCIR CrossLink organisers (see Chapter 7). It causes the ground truth to get larger, which has two consequences: (1) It increases the measured (accuracy of) precision, as many relevant links are often omitted in the ground truth. (2) It is more difficult to achieve high recall, as there are too many links.

Even when ground truths are combined in this way, it is expected that one of the issues of this evaluation approach can still be the perceived subjectivity of the linking task. As a result, in addition to carrying out the evaluation using this standard approach, we find it essential to estimate the agreement between annotators and see how the measured precision and recall characteristics compare with link structures created by different groups of people. We will address this issue in Section 6.4.3.

6.4 Results

6.4.1 Experimental setup

The experiment was carried out for two language pairs: Spanish to English and Czech to English. We will denote the source language L_{source} and the target language L_{target}. The input for the different CLLD methods are two document sets:

- Let $SOURCE_{L_{source}}$ be the set of topic documents selected as pages that contain a Wikipedia link between different language versions. In our case, 100 topics (pages) were sampled and selected as link sources. Please note that hundreds of links can be generated from each topic document.

- Let $TARGET_{L_{target}}$ be the collection of documents in the target language from which the link targets are selected. In our case, this collection contains all (3.8 million) Wikipedia pages in English.

The output of the method is a set (ranked list) $LIST_{result} = \langle TARGET_{L_{target}}, score \rangle$. To establish the ground truth we define:

- Let ρ be the mapping from documents in the source language to their target language versions $\rho : D_{L_{source}} \rightarrow D_{L_{target}}$.

- Let $SOURCE_{L_{target}}$ be the set of topic documents mapped to the target language $SOURCE_{L_{target}} = \rho SOURCE_{L_{source}}$.

- Let α, β be the mappings from documents to the other documents they link to in the source and target language respectively $\alpha : D_{L_{source}} \rightarrow D_{L_{source}}$, $\beta : D_{L_{target}} \rightarrow D_{L_{target}}$.

then we define the ground truth (GT) as the union of ground truths for different language versions, in this experiment we define it as the union of ground truths for the source and target language.

$$GT = \alpha(SOURCE_{L_{source}}) \cup \beta(SOURCE_{L_{target}})$$

A given generated item $\langle d, score \rangle \in LIST_{result}$ is evaluated as a hit if and only if $d \in GT$.

6.4.2 Methods evaluation

To investigate the performance of the first part of CLLD, i.e. the cross-language step carried out by CL-ESA, we have analysed how well the system finds for a given topic document in the source language the duplicate document in the target language. In this step, the system takes a document in the source language, and selects from the 3.8 million large document set in the target language the documents with the highest similarity. We then check, if a duplicate

document ($d = \rho d_{source}$) appears among the top k retrieved documents. The experiment is repeated for all examples in $SOURCE_{L_{source}}$ and the results are then averaged (Figure 6.5). The graph suggests that the method performs well, as the document often appears among the first few results. In about 65% of cases, the document is found among the first 50 retrieved items. We believe that if the set of candidates (controlled by the t parameter) contains this document, the CLLD method is likely to produce better results, this is especially true for the CL-ESA2Links method.

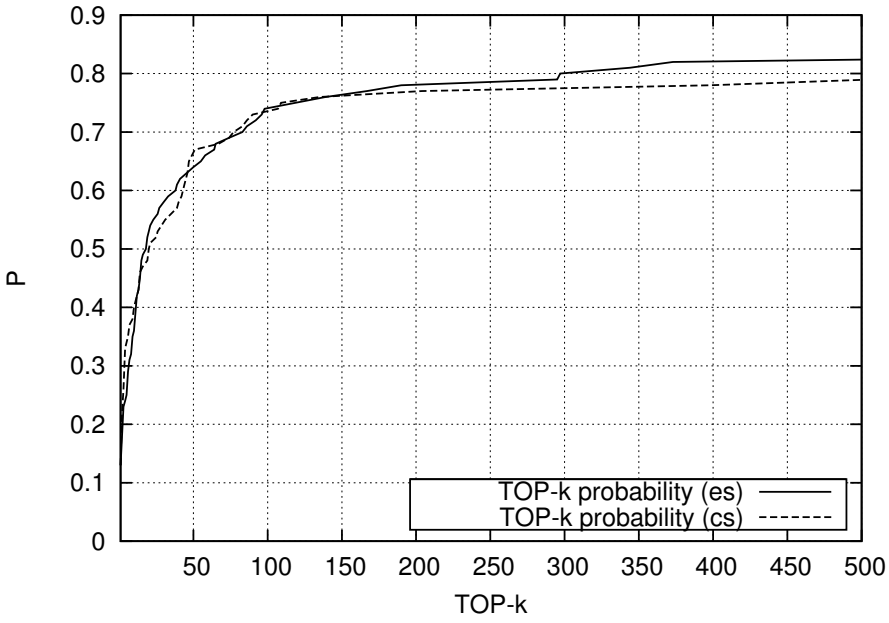


Figure 6.5: The probability (y-axis) of finding the target language version of a given source language document using CL-ESA in the top k retrieved documents (x-axis). Drawn as a cumulative distribution function.

The overall results for all the methods are presented in Figure 6.6. We have

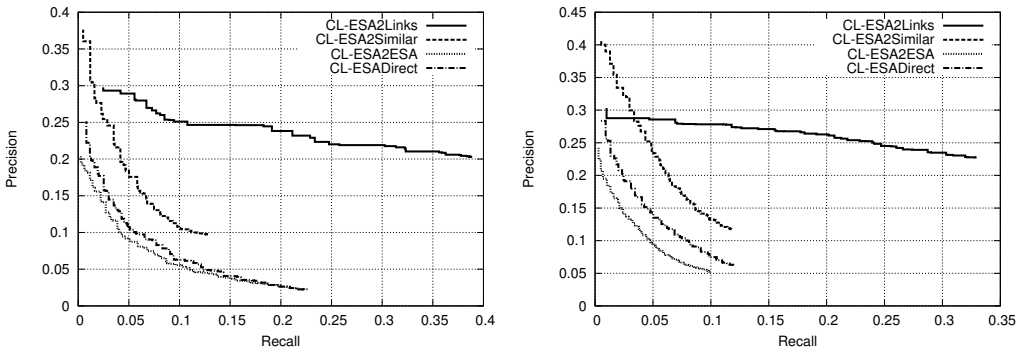


Figure 6.6: The precision (y-axis)/recall (x-axis) graphs for Spanish to English (left) and Czech to English (right) CLLD methods.

experimentally set $t = 10$ for Spanish to English and $t = 3$ for Czech to English CLLD. CL-ESA2Links performed in the experiments the best achieving 0.2 precision at 0.3 recall. CL-ESA2Similar performed the best out of the purely content-based methods.

Though the precision/recall might seem quite low, a number of things should be taken into account:

- A significant number of potentially useful links is still missing in our ground truth, because people typically do not intend to link all relevant information. As a result, many potentially useful connections are not explicitly present in Wikipedia (see Section 4.5.1). The problem can be partly mitigated by taking a union of the ground truths from more language versions. Another approach is to measure the agreement instead of precision/recall characteristics (see Section 6.4.3).

- A significant number of links in Wikipedia are conceptual links. These

links do not express a particularly strong relationship at the article level. This makes it very difficult for the pure-content based methods to find them, which results in low recall. It seems that CL-ESA2Links is the only method that does not suffer from this issue.

- The experiment settings make it hard for the methods to achieve high precision/recall performance. The $TARGET_{L_{target}}$ set contains 3.8 million articles, out of which, the methods are supposed to identify just a small subset of target documents. More precisely, in Spanish to English CLLD, our ground truth contains on average 341 target documents with standard deviation 293, in Czech to English, it contains on average 382 target documents with standard deviation 292.

6.4.3 Measuring the agreement

To assess the subjectivity of the link discovery task and to investigate the reliability of the acquired ground truth, we have compared the link structures from different language version of Wikipedia. We have iterated over the set of topics from $SOURCE_{L_{source}}$ and recorded for each document in $TARGET_{L_{target}}$ in each step if it is a valid link target (yes - Y) or if it is not a valid link target (no - N) for the given source document in each language, thus measuring the agreement between the link structures in different languages. The results are presented in Table 6.1.

Spanish vs English			
	Y_{en}	N_{en}	N/A_{en}
Y_{es}	5,563	10,201	3,934
N_{es}	15,715	539,299,641	99,191,766
N/A_{es}	5781	321,326,145	0
Czech vs English			
	Y_{en}	N_{en}	N/A_{en}
Y_{cz}	4,308	8,738	2,194
N_{cz}	12,961	392,411,445	7,501,806
N/A_{cz}	9,790	356,532,740	0

Table 6.1: The agreement of Spanish and English Wikipedia and Czech and English Wikipedia on their link structures calculated and summed for all pages in $SOURCE_{en}$. Y — indicates yes, N — no, N/A — not available/no decision

As demonstrated in Figure 6.7, a subset of Wikipedia article pairs cannot be mapped to other language versions (because at least one of the articles in the pair does not have an equivalent in the other language). These links were classified as no decision/not available (N/A). The mappable pairs were classified in a standard way according to their appearance in the link graphs of the language versions. Only these links are taken into account when measuring the agreement.

A common way to assess inter-annotator agreement between two raters in information retrieval is using Cohen's Kappa introduced in Section 3.3.2 and calculated as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{6.1}$$

where $Pr(a)$ is the relative observed frequency of agreement and $Pr(e)$ is

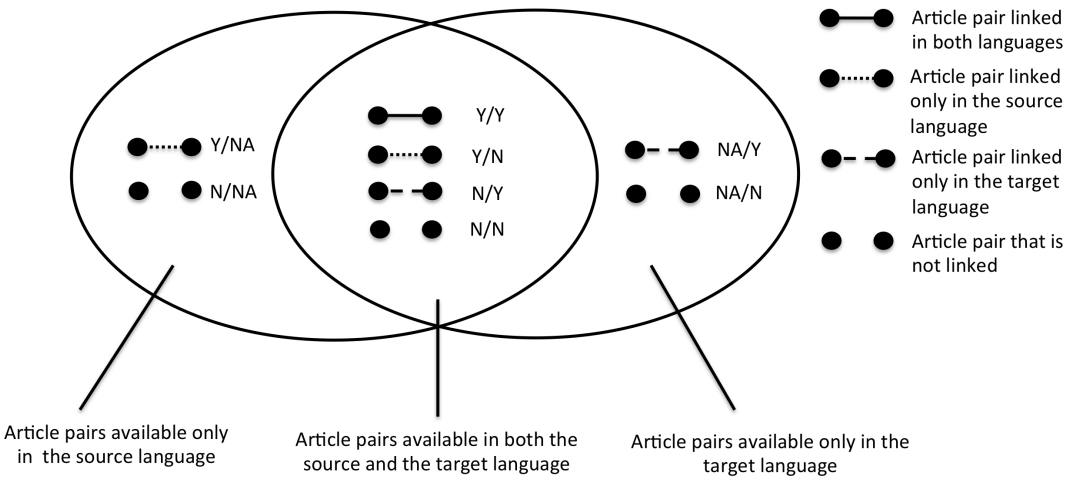


Figure 6.7: Individual cases of agreement, disagreement and no decision on linking a Wikipedia article pair in two language versions of Wikipedia link graphs. Y, N and NA indicate if a link connects an article pair, does not connect it or if an article pair does not exist in a given language version of Wikipedia respectively. The cases correspond to the combinations of connections of article pairs in the source and the target Wikipedia language versions.

the hypothetical probability of chance agreement. $Pr(a)$ is typically calculated as $\frac{|Y,Y|+|N,N|}{|Y,Y|+|Y,N|+|N,Y|+|N,N|}$. Since there is a strong agreement on the negative decisions, the probability will be close to 1. If we ignore the $|N,N|$ cases, which do not carry any useful information, the formula looks as follows:

$$Pr(a) = \frac{|Y,Y|}{|Y,Y| + |Y,N| + |N,Y|}. \tag{6.2}$$

The probability of a random agreement is extremely low, because the probability of a link connecting any two pages is approximately:[4]

[4]Following the official Wikipedia statistics. Though different language versions have different p_{link}, the differences do not affect the results.

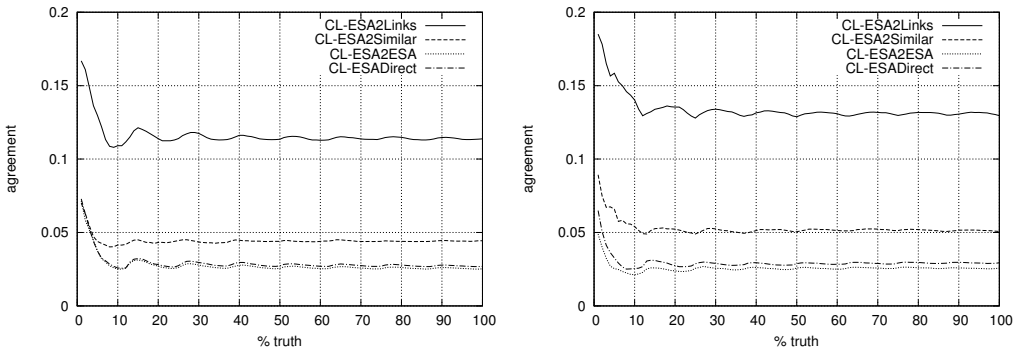


Figure 6.8: The agreements of the Spanish to English (left) and Czech to English (right) CLLD methods with $GT_{es,en}$ and $GT_{cz,en}$ respectively. The y-axis shows the agreement strength and the x-axis the number of generated examples as a fraction of the number of examples in ground truth.

$$p_{link} = \frac{|links|}{|pages|^2} = \frac{78.3M}{3.2M^2} = 0.000007648. \tag{6.3}$$

Thus, the hypothetical number of items appearing in the Y, Y class by chance is $p_{link}^2.(|Y,Y| + |Y,N| + |N,Y| + |N,N|)$. This formula estimates the number of agreements achieved by chance. In our case the value is much smaller than one,[5] hence $Pr(e)$ is close to 0. Therefore, we can calculate the agreement for English and Spanish as:

$$\kappa_{en,es} = \frac{5,563}{31,479} = 0.177. \tag{6.4}$$

The agreement for Czech and English is:

[5]Meaning that there is a low probability that even a single agreement would be observed by chance.

$$\kappa_{en,cz} = \frac{4,308}{26,007} = 0.166. \tag{6.5}$$

The value indicates a relatively low inter-annotator agreement. We believe that the fact that such a low agreement has been measured is very interesting, particularly because the link structure in Wikipedia is a result of a collaborative effort of many contributors. Therefore, we would expect that even lower agreement might be experienced in other types of text collections.

Motivated by the previous findings, we have calculated the agreement between the output of our method and the link graphs present in different language versions of Wikipedia. We were especially interested to find out if the agreement is significantly different from the agreement measured between different language versions of Wikipedia. We have generated by our CLLD methods 100% of $|GT|$ links for every orphan document in $SOURCE_{L_{source}}$, i.e. if a particular document is linked in Wikipedia to 57 documents, we generate 57 links. We have then measured the agreement for each topic document and averaged the agreement values. The results of the experiment for Spanish to English and Czech to English CLLD are shown in Figure 6.8. They suggest that CL-ESA2Links achieved a level of agreement comparable to that of human annotators. A very reasonable level of agreement has also been measured for CL-ESA2Similar, especially for the first 10% of the generated links. CL-ESADirect and CL-ESA2ESA exhibit a lower level of agreement.

6.5 Summary of contribution

The aim of this chapter was to investigate how can we generate links between semantically related texts written in different languages, to explore issues related to the evaluation of CLLD methods and to discuss how the links produced by CLLD methods compare to those produced by different language communities. The main contributions of this chapter are the design of the new CLLD methods and the new insights into the evaluation of CLLD systems.

More specifically, we have presented and evaluated four CLLD methods, one link-based and three text-based. We have also been the first to apply ESA (and CL-ESA) to the link discovery task. The evaluation results suggest that methods that are aware of the link graph in the target language have the potential to achieve slightly better results than text-based methods that identify links in the target language by calculating semantic similarity. However, as the former methods cannot be applied in all document collections, the latter methods are valuable.

Although it might seem at first sight that CLLD methods do not provide very high precision and recall, we have demonstrated this is largely the artefact of the unreliability/disparity of the available ground truths. This motivated us to measure the agreement between ground truths extracted from different language versions of Wikipedia, assuming that content is linked due to the

relatedness of semantics and the resulting link graph should therefore be highly correlated regardless of the language used to express the semantics. The results show that the agreement between link graphs of different language versions of Wikipedia is surprisingly low. This explains why one needs to be very careful in interpreting the results of link discovery methods when they are evaluated against a ground truth extracted from Wikipedia. It needs to be understood that precision can be artificially decreased due to many valid links missing in the ground truth. On the other hand, combining ground truths from multiple language versions might result in an artificial inability of the methods to achieve higher levels of recall. A solution here might be the use of graded relevance (see Chapter 3), which we will further discuss in the next chapter. Finally, if the results produced by our CLLD methods are taken as a link structure of another language, comparable agreement to that of the link structures produced by different language communities can be seen. This might suggest that the performance achieved by our CLLD methods is actually fairly close to that which can be achieved by independent communities of human annotators.

In the next chapter, we will further address the same research questions, however, focusing on noun phrase-to-document CLLD.

Chapter 7

Noun Phrase-to-Document Cross-Lingual Link Discovery (CLLD) in Wikipedia

The previous chapter explored how link discovery can be extended to deal with multilingual text collections. This chapter further builds on this research work. Its aim is to investigate how to develop robust noun phrase-to-document CLLD systems tailored to a specific application domain. The chapter is based on the research and the results we achieved at two consecutive link discovery evaluation conferences organised by the National Institute of Informatics Testbeds and Community for Information Access Research (NTCIR) project. These two CLLD evaluation forums were called CrossLink-1 and CrossLink-2 and were associated with NTCIR-9 (2011) and NTCIR-10 (2013) respectively. NTCIR is a major international forum (similar to TREC) of evaluation tracks

designed to enhance research in Information Access (IA) technologies including information retrieval, question answering, text summarisation and extraction (*About NTCIR*, 2014). The NTCIR conferences take place every 18 months in Tokyo, Japan, but participants are expected to work on their methods over the 18 months prior to the conference.

The CrossLink task (Cross-Lingual Link Discovery — CLLD) aims to automatically find anchors and links from these anchors to appropriate Wikipedia documents across languages. The task is concerned with access between English and Asian languages, particularly Chinese, Japanese and Korean. As highlighted by the task organisers, the CrossLink task is not directly related to traditional cross-lingual information retrieval (CLIR), because CLIR can be viewed as a process of creating a virtual link between the provided cross-lingual query and the retrieved documents; but CLLD actively recommends a set of meaningful anchors in the source document and uses them as queries with the contextual information from the text to establish links with documents in other languages. The CrossLink task is closely related to the work we have presented in the previous chapter and further addresses RQ 3 and RQ 4. Additionally, the chapter aims to fulfill Goal 1:

Goal 1: *Design new link discovery methods and evaluate them under the umbrella of an international evaluation conference, such as NTCIR, in direct competition with other research teams.*

NTCIR CrossLink provides an excellent opportunity and framework for experimenting with new noun phrase-to-document CLLD methods and their evaluation in direct competition with methods designed by other researchers.

In the following text, we will often use the terms *anchor, concept, term, link, sense, target, outlink and Wikipedia version* in the following way. By *term*, we understand any textual fragment (typically a noun phrase) that can be potentially used as the (clickable) body of a hypertext *link*. By *anchor*, we understand an actual instance of a term used as the body of a hypertext link. We will refer to instances of the Wikipedia collection written in different languages as *Wikipedia versions*. Every Wikipedia page describes a *concept*. The name of the described concept is usually provided as the title of the Wikipedia page. Though concepts are, in principle, language independent, we will refer to the page an ordinary monolingual link points to as the *concept* and to an equivalent page in another language as the *equivalent concept*. A *link* is consequently defined by an anchor-concept pair and a *cross-language link* by an anchor-equivalent concept pair. Alternatively, the CrossLink terminology uses the term *target* to refer to the concept linked by an anchor and the term *outlink* to refer to a link from a particular concept. We can say that every anchor in a Wikipedia version links to a concept in the same Wikipedia version. A concept in a Wikipedia version can have an equivalent concept in another Wikipedia version. A concept can be linked to from many (synonymous)

anchors. Different anchors can use the same term to link to different concepts (we say the term can refer to multiple *senses*).

Based on these definitions, the *CrossLink task* can be described as follows: given a new concept *(orphan document)*[1] in the source language, the goal is to identify a ranked list of suitable anchors in the orphan document and link them to relevant concepts (*targets*) in the destination language version of Wikipedia.

While Wikipedia has been used as a dataset in the previous chapters of the thesis, its use has been quite different from the one in CrossLink. In the previous chapters, we used Wikipedia only as a set of documents with links. Our methods aimed at finding pairs of documents that can be connected using a link, while not relying on information that is specific to the Wikipedia collection, such as the information about the:

- Correspondence of anchors to the names of Wikipedia pages.
- Disambiguation of noun phrases that are used as link anchors.
- The set of all possible noun-phrases from which links can originate.

As the problem we address here can be classified as noun phrase-to-document link discovery and is performed specifically for Wikipedia (*wikification*, see Section 3.2.1.2), some of our methods will rely on this information (in particular those presented in Section 7.2). The focus on noun phrase to document link

[1]The term orphan document is used by the task organisers to refer to a new Wikipedia page without any link markup.

discovery and the domain specificity of the task distinguishes this chapter from the previous one.

The methods we applied in CrossLink-2 follow quite different strategies than the methods we used in CrossLink-1 (Knoth et al., 2011). While in CrossLink-1 we approached the problem as a similarity search task, in CrossLink-2 we saw it rather as a disambiguation and ranking problem. Both approaches have advantages and disadvantages. In CrossLink-1, we designed methods that are quite general and extensible in their ability to be applied to interlinking in non-Wikipedia contexts (e.g. newspapers, blogs or books). On the other hand, the CrossLink-2 methods are tailored to the Wiki (or even Wikipedia) environment (and thus they are also closer to the methods of most other CrossLink participants). These methods consequently achieve better results on the CrossLink dataset.

Evaluation conferences, including NTCIR and TREC, use certain terminology to organise the evaluation process. We think the knowledge of this terminology is essential for the understanding of the chapter. A *topic* represents a specific information need selected to be used in the evaluation. In CrossLink, a topic is synonymous to one orphan document selected by the task organisers to be interlinked with the rest of the collection using the CLLD methods. A *test collection* to be used in the evaluation will typically contain more than one topic (25 in CrossLink). The term *run* refers to the output file, conforming to

the format specified by the task organisers, containing the calculated results for all topics in the test collection using one method. The term *submission* refers to multiple runs submitted by a *team* to the evaluation. Please note that the term *team* is used in the NTCIR terminology to refer to one or more authors of the CrossLink submission.

The rest of the chapter is organised as follows. Section 7.1 presents new methods we submitted for evaluation to NTCIR-9 CrossLink, while Section 7.2 presents the methods we submitted to NTCIR-10 CrosLink-2. Both our submissions are denoted as "Team KMI" (Knowledge Media institute) in the evaluation overview papers of CrossLink (Tang, Geva, Trotman, Xu and Itakura, 2011) and CrossLink-2 (Tang et al., 2013). These papers might serve as a valuable reference to this chapter as they contain the evaluation graphs and tables for all subtasks and teams and also provide an overview of the results achieved by different participating teams. A summary of contribution of this chapter is presented in Section 7.4.

7.1 KMI @ NTCIR-9 CrossLink: CLLD in Wikipedia as a similarity search problem

This section describes the methods used in our submission to NTCIR-9 CrossLink. We submitted four runs for link discovery from English to Chinese; however,

the presented methods are applicable also to other language combinations. The CLLD problem is approached as a similarity search tasks using ESA (see Chapter 6) as a criterion of similarity. Three of our runs are based on exploiting the cross-lingual mapping between different language versions of Wikipedia articles. This mapping is explicitly present in Wikipedia. In the fourth run, we assume information about the mapping is not available and we use CL-ESA instead (in a similar way as in Chapter 6). Our methods achieved encouraging results and we describe in detail how their performance can be further improved. Finally, we discuss two important issues in link discovery: the evaluation methodology and the applicability of the developed methods across different textual collections.

NTCIR-9 CrossLink was the first evaluation forum to stimulate the development and compare the performance of CLLD systems for Wikipedia. The methods submitted by different teams typically build on successful monolingual systems and solutions developed and previously tested at INEX: Link The Wiki Track evaluations adapted to the multilingual environment. The most common ways of dealing with the issue of multilinguality are (a) using the manually defined mappings between equivalent Wikipedia pages or (b) using machine translation. This is why in one of our runs, we have explored the possibility of applying CL-ESA to this problem.

In Section 7.1.1, we describe our methods and provide the description of

our runs. Section 7.1.2 presents the achieved results providing a comparison with other participants. Section 7.1.3 discusses the approach and Section 7.1.4 concludes the NTCIR-9 experiments.

7.1.1 Link discovery methods

The CLLD methods we have developed operate in three phases: *target discovery*, *anchor detection* and *link ranking*, as demonstrated in Figure 7.1. In the first phase, we take the orphan document (topic) in the original language and try to find other documents in the target language that could be considered suitable link targets, using semantic similarity as a criterion. In the second step, we take the list of candidate targets and try to detect for each of them a suitable anchor in the orphan document. In the third phase, we describe each anchor using a set of features and perform link ranking using a Support Vector Machine (SVM) classifier. The following subsections describe them in more detail.

7.1.1.1 Target discovery

In the target discovery phase we take as an input a new "orphan" document (i.e. a document that is not linked to other documents) written in the source language and we automatically generate a list of potential target documents. In this phase, the system works at the granularity of the whole documents.

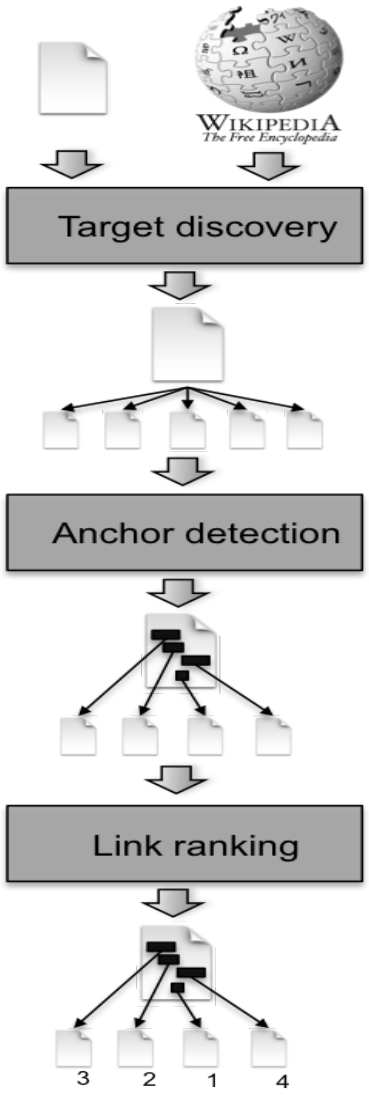


Figure 7.1: Cross-Lingual Link Discovery process

We apply two different approaches to accomplish this task. The first approach is based on the application of ESA in combination with the existing link structure of Wikipedia, and we will call it *ESA2Links*. As the name suggests, the approach is similar to the CL-ESA2Links method described in Chapter 6, with the difference that we apply only monolingual ESA in the first step, instead of CL-ESA. The second approach utilises the information about Wiki page titles, and we will call it *Terminology*. Both approaches can be combined or used separately.

The *ESA2Links* method works in two steps. In the first step, an ESA vector is calculated for each document in the document collection. This results in obtaining a weighted vector of Wikipedia concepts for each document in the source language. The cardinality of the vector is given by the number of concepts (i.e. pages) in the source language version of Wikipedia (about 3.8 million for English). The same procedure is applied on the orphan document. Similarity between the resulting ESA vectors is then calculated and the k most similar pages are identified. In our runs we use $k = 1,000$. This value of k was experimentally selected, as it showed good performance when evaluated against the Wikipedia ground truth.

In the second step, the k most similar documents to the orphan document are taken as a seed and are used to discover documents that are suitable link targets. In Chapter 6, we have described and evaluated four alternative

approaches to target discovery after ESA is applied. The Link approach, which produced the best results has been used. As this approach requires access to the link structure in the document collection, please see Chapter 6 for alternatives that do not have this requirement. After generating the seed documents, the method extracts all links in the form $[anchor, pageID]$ present in those seed documents, where $pageID$ is the Wiki identifier of the anchor destination. Using the cross-lingual mapping between Wikipedia pages, the $pageID$, describing a page in the source language, is mapped to an appropriate ID describing the same page in the target language. If the mapping is not explicitly specified in Wikipedia, the link is discarded. The resulting set of pairs represents the set of candidate targets.

The *Terminology* approach is much simpler than the previous one and can be considered the baseline approach. The method exploits the title information of Wiki articles and the cross-lingual mapping between Wikipedia articles. The method recommends as targets all pairs $[pageTitle, pageID]$ in the whole Wikipedia for which there exists an explicit cross-lingual mapping between the source and the target language version of the page, i.e. the resulting set of targets will be always the same regardless of the orphan document. It is up to the next phase to filter down the list of targets to those that are suitable.

7.1.1.2 Anchor detection

In the anchor detection phase, we take as an input the set of targets and try to detect suitable anchors for them in the orphan document. The procedure is quite simple: We iterate through the set of target documents and we try to find a suitable anchor text in the orphan document given the target document title. If no anchor is discovered, the link is discarded.

The simplicity of this phase is very much given by the fact that the methods are tailored for Wikipedia. In Wikipedia, each page is characterised by a title. In addition, the anchor texts in Wikipedia are typically identical to the name of the title of the page which describes a given concept or are variations of the title which can easily be extracted from the collection. This is not the case in general (non-Wiki style) text collections where this step is significantly more challenging given the variability of link types (see Chapter 4).

7.1.1.3 Link ranking

In the link ranking phase, we take the list of links in the form $[anchor, targetID]$, where *anchor* represents the specific text in the orphan document and the $targetID$ is the Wiki *page ID* of the target page in the target language, and we rank the links according to their importance defined as the confidence of the ranking system.

The approach we are using to generate our runs is based on machine learn-

ing. Each link is first described and modelled by a set of features (occurrence, generality and link frequency are inspired by Milne and Witten (2008)). The features are represented as a vector assuming their mutual independence.

- *ESA similarity* is a real number between 0 and 1, which expresses the similarity of text. Three different features were included:
  - Similarity of the link text to the target document text.
  - Similarity of the link text to the target document title.
  - Similarity of the input document text to the target document text.
- *Generality* is a measure expressing how general a given topic is. It is an integer number between 0 and 16 defined as the minimum depth at which the topic is located in Wikipedia's category tree.
- *Link frequency* is a measure expressing how many times a particular keyword occurs as a link (or more precisely as an anchor) in the whole document collection.
- *Occurrence of the link text in the input document* is a relative measure of the first, last and current occurrence of the link text in the input document, and the difference between its first and last occurrence.

When the features are encoded, we train a Support Vector Machine (SVM). In our experiments, the system was trained on the examples associated to the

three topic documents provided by the NTCIR CrossLink organisers. Negative examples were acquired by running the *ESA2Links* and anchor detection method described above, and by filtering out the positive examples provided by the organisers. In the testing phase the SVM classifier is used to decide whether a link should be included. Given the low number of training examples, we expect the SVM to have a relatively low recall, but high precision. The confidence value of the SVM, which characterises the distance from the decision hyperplane, is used to select the best candidates. The candidates are then ranked according to their semantic similarity to the orphan document.

7.1.1.4 Cross-lingual discovery

We submitted four runs out of which three use the explicit information about cross-lingual mapping between Wiki pages. This makes the methods more difficult to reuse in other contexts. As a result, we have also tested in one of our runs a more challenging setting in which we utilise Cross-Lingual Explicit Semantic Analysis (CL-ESA) to discover an equivalent page in the target language (Chinese) for a page in the source language (English). The method is based on the mapping of the ESA conceptual space between the two languages. In our runs, we refer to this approach as *ESA discovery*.

The most semantically similar target language document to the orphan document is considered by the method as a suitable candidate. To identify

such a document, cosine similarity is calculated between the ESA vector of the source document with the ESA vectors of other documents in the target document collection in the same way as decribed in Section 6.1.1.

7.1.1.5 KMI runs

We have submitted four runs for NTCIR-9 CrossLink for English to Chinese. While the methods are applicable to other language combinations, we have tested them on Chinese only.

Run 1: *KMI_SVM_ESA_TERMDB* - combines *ESA2Links* with *Terminology*

Run 2: *KMI_SVM_ESA* - applies *ESA2Links* for target discovery.

Run 3: *KMI_SVM_TERMDB* - uses *Terminology* only for target discovery.

Run 4: *KMI_ESA_SVM_ESAdiscovery* - uses *ESA2Links* for target discovery and *ESA discovery* for the cross-language step.

7.1.2 Experiments

7.1.2.1 Evaluation methodology

All links and supporting information were cleared from the English articles used in the evaluation. The remaining link structure has been kept. The methods have been evaluated at different granularity levels: anchor-to-file (A2F)

and file-to-file (F2F). There were two evaluation modes: automatic and manual (see Section 3.3.2).

- *Automatic assessment* — the ground truth is derived automatically from the existing link structure of Wikipedia.
- *Manual assessment* — all anchors and targets are pooled and the evaluation is carried out by a human assessor.

Precision-at-N (P@N), *R-Prec*, and *Mean Average Precision (MAP)* were used as the main metrics to evaluate the performance of the CLLD methods. More information about the ground truth, the evaluation setup and a detailed description of the evaluation measures can be found in the overview paper Tang, Geva, Trotman, Xu and Itakura (2011).

7.1.2.2 Evaluation

Runs were evaluated on 25 topic documents. For each topic document there was a maximum limit of 250 anchors that could be extracted, each of which could point to the maximum of 5 different target documents (i.e. the maximum number of generated links per topic document was set to 1,250 by the task organisers). All four KMI runs were submitted for English to Chinese. The F2F performance of the KMI methods with Wikipedia ground truth is shown in Figure 7.2. There is no A2F evaluation with Wikipedia ground truth as such evaluation would be difficult for a number of reasons: "An anchor

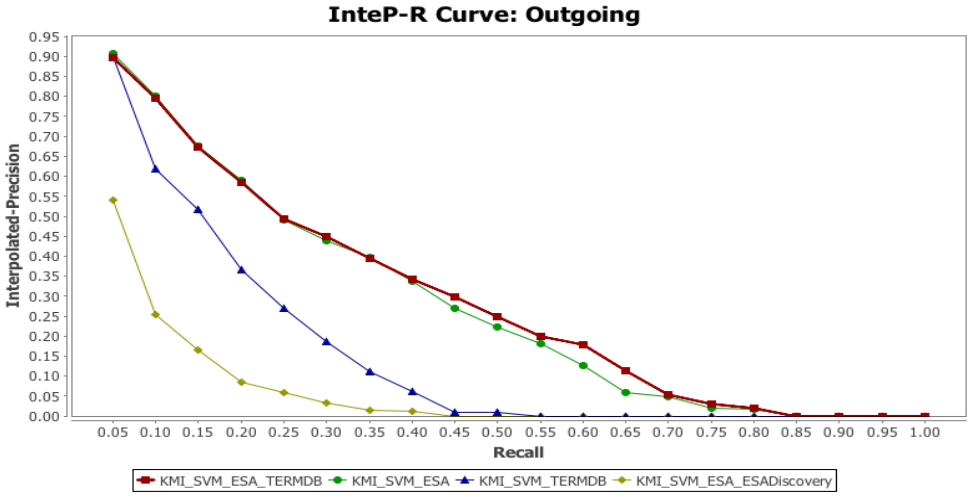


Figure 7.2: F2F performance of the KMI runs using Wikipedia ground truth.

can occur multiple times in a document in subtly different linguistic forms. It is unreasonable to score multiple identical links and also unreasonable not to score different linguistic variants. The best approach to measuring this imprecision is unclear and has been studied at the INEX Link Discovery Track where it changed from year to year" Tang, Geva, Trotman, Xu and Itakura (2011). Figure 7.3 and Figure 7.4 show the performance of the presented methods when manual assessment has been used for both F2F and A2F granularity levels. The results for all experiments are summarised in Table 7.3.

7.1.2.3 Comparison of the runs performance

Overall, we can see that the *KMI_SVM_ESA_TERMDB* method achieved the best results in terms of MAP and R-Prec in all evaluations. Very similar results have been achieved by the *KMI_SVM_ESA* method showing that the use

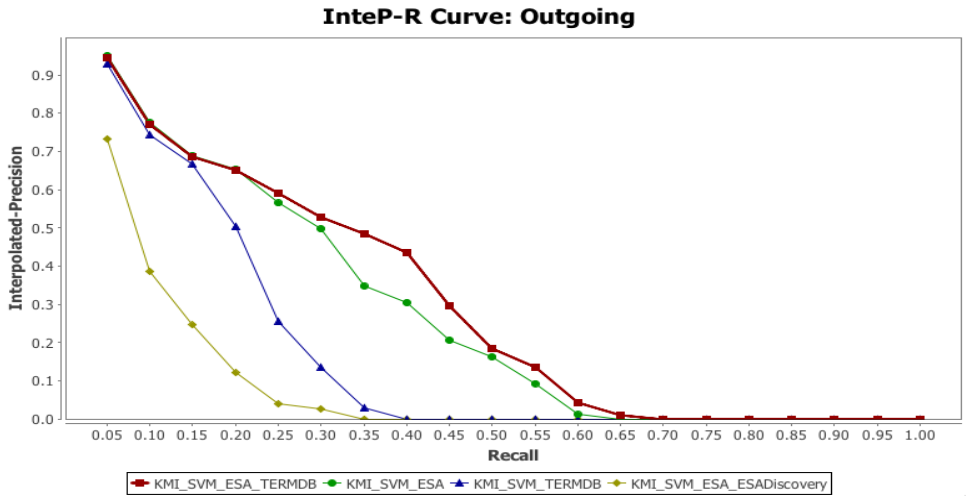


Figure 7.3: F2F performance of the KMI runs using manual assessment.

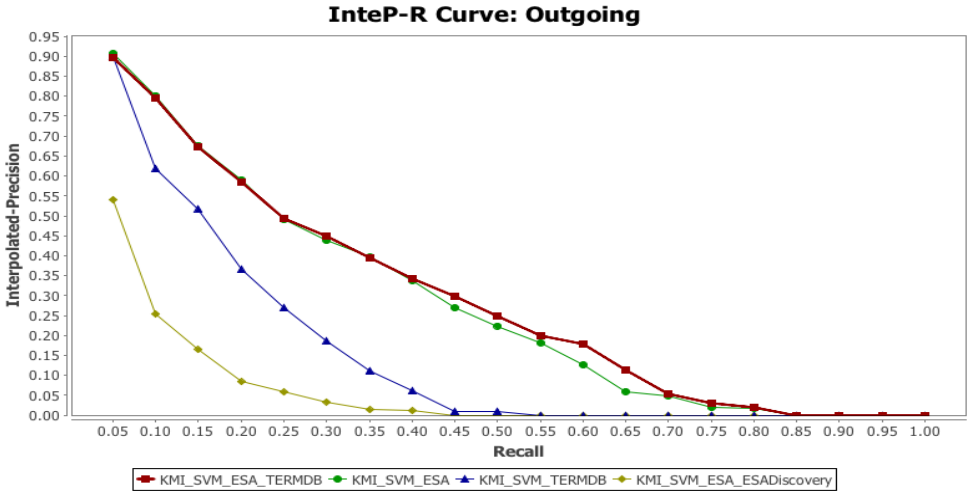


Figure 7.4: A2F performance of the KMI runs using manual assessment.

Run ID	MAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
F2F performance with Wikipedia ground truth								
KMI_SVM_ESA_TERMDB	**0.260**	**0.345**	0.712	**0.664**	0.530	0.491	**0.434**	**0.166**
KMI_SVM_ESA	0.251	0.338	**0.728**	**0.664**	**0.540**	**0.493**	0.430	0.153
KMI_SVM_TERMDB	0.127	0.211	0.624	0.552	0.454	0.383	0.302	0.078
KMI_ESA_SVM_ESADiscovery	0.059	0.148	0.264	0.240	0.186	0.165	0.138	0.044
F2F performance with manual assessment results								
KMI_SVM_ESA_TERMDB	**0.258**	**0.393**	0.720	**0.728**	**0.684**	0.648	0.604	**0.358**
KMI_SVM_ESA	0.231	0.344	0.728	0.720	0.678	**0.668**	**0.615**	0.306
KMI_SVM_TERMDB	0.133	0.192	**0.752**	0.692	0.636	0.613	0.561	0.178
KMI_ESA_SVM_ESADiscovery	0.054	0.132	0.464	0.388	0.348	0.321	0.283	0.119
A2F performance with manual assessment results								
KMI_SVM_ESA_TERMDB	**0.097**	**0.114**	**0.368**	**0.368**	**0.330**	0.303	0.269	**0.142**
KMI_SVM_ESA	0.080	0.092	0.360	0.364	**0.330**	0.299	0.260	0.113
KMI_SVM_TERMDB	0.070	0.075	0.376	**0.368**	0.324	**0.316**	**0.297**	0.096
KMI_ESA_SVM_ESADiscovery	0.014	0.035	0.088	0.108	0.110	0.108	0.090	0.045

Table 7.1: Performance of the KMI methods

of the terminology dictionary in the target discovery step helps only moderately. The *KMI_SVM_TERMDB* method produced in most cases substantially worse results than the two methods that used the *ESA2Links* approach. This shows that combining semantic similarity with the information about existing Wikipedia links provides valuable information.

It is not surprising that the *KMI_ESA_SVM_ESADiscovery* method produced on this dataset worse results than the other methods as it is the only method that makes use of the explicit (manually created) cross-language mapping between different language versions of Wikipedia articles. On the other hand, this method is more generally applicable than the other methods.

Team	Anchor source	Anchor ranking	Disambiguation	Translation	Key features
HITS	Wiki-Titles	Prior probability	Maximum edge weighted clique	Triangulation	Multilingual concept repository Anchor disambiguation
UKP	Noun phrases, Named entities, Wiki-Anchors, Wiki-Titles, N-Grams	BM25, Anchor probability, Anchor strength	N/A	Triangulation, Dictionary, Machine translation	Several anchor selection methods, Several anchor text translation methods
QUT	Wiki-Links	Link Probability, Page Name Matching, CLIR	N/A	Triangulation, Machine translation	Mostly link prior probability
KMI	Wiki-Titles	Link Ranking	N/A	Similar concept clustering	Target documents first Ranking with SVM

Figure 7.5: Comparison of CLLD methods of four participating teams with the overall best results. The table is taken from (Tang et al., 2014).

7.1.2.4 Performance comparison with other teams

The KMI methods scored first in Precision-at-5 in the A2F manual assessment and third in terms of R-Prec. Our methods were also second in F2F manual assessment in terms of MAP and R-Prec and third in terms of Precision-at-5. Our system ranked third in F2F Wikipedia ground-truth evaluation in terms of all MAP, R-Prec and Precision-at-5. Figure 7.5 provides a methods comparison of four participating teams with the overall best results. In total, 57 runs from 11 teams were evaluated. For more information on this topic, see (Tang et al., 2014).

7.1.2.5 Unique relevant links

The CrossLink organisers decided this year to also compare the systems based on the number of unique relevant links the individual systems have contributed (Tang, Geva, Trotman, Xu and Itakura, 2011). Our methods ranked in the comparison third for Wikipedia ground-truth (27 unique relevant links) and

second for manual assessment (152 unique relevant links). However, we believe the results of this comparison should be interpreted very carefully because:

- This evaluation metric was not known to the participants prior to the submission and therefore the system parameters were not optimised to achieve high results in this evaluation.

- The results that are being compared are the number of unique links provided by the runs of different teams and therefore teams that have submitted less runs than the others are at a disadvantage.

- The comparison puts systems that have not generated all allowed 1,250 links per topic at a disadvantage. For example, a run producing high precision results can receive low score according to this measure in case it does not decide to generate all 1,250 links (a constant defined by the task organisers). Since this evaluation measure does not take into account the assigned rank to a particular link, a system that has generated, for example, 200 good links will receive a lower score than a system that has generated first 1,000 links wrongly and the last 250 links are correct.

- It should be expected that the number of unique relevant links generated can differ significantly based on the selection and variation of parameters of different systems. Therefore, in the future, such an evaluation should be carried out by taking into account the sensitivity of the sys-

tems to parameters, and the trade-off between unique relevant links and precision/recall characteristics.

7.1.2.6 How can the performance be improved?

There is a number of ways in which our methods could be improved and optimised for better performance. We see the main possibilities in:

Extending the set of training examples — the link ranking system (step 3) has been trained on a very limited number of examples. These examples included links relevant to only three topic documents provided by the organisers (i.e. Australia, Femme fatale and Martial arts). It is therefore reasonable to assume that just a moderately larger training data could increase the ranker performance.

Extending the methods to enable linking all articles — The three best performing methods we have presented rely on the existence of cross-lingual links in the Wikipedia collection. Our experiments show that for a large proportion of Chinese articles the mapping to English is missing. Therefore, our methods could be improved if this information was present in Wikipedia or by using methods that can detect different language versions for a Wikipedia article. Such a method was, for example, presented in Sorg and Cimiano (2008*b*).

Dimensionality of the ESA vector — to be able to run the methods quickly on our machines we decided to represent each document using only the best 100

ESA dimensions. The other dimensions of the vector were set to zero. While our experiments show that preserving only the best 100 dimensions strongly correlates (0.825 Spearman's rank correlation) with the results produced with 10,000 dimensions, preserving 2,500 dimensions would result in an almost perfect correlation of 0.98. We can assume that this could slightly improve the results of the three best performing methods and significantly improve the results of the method that makes use of cross-language discovery using ESA (Run 4).

The cross-language discovery step — we analysed the method relying on cross-language discovery of Wikipedia articles (Run 4), in particular the step in which the system takes an English article and tries to automatically discover the version of the same article in the target language. This task is difficult as the system has to select the correct article given the set of all articles in the target language. For Chinese, this amounts to $318,736$ documents. Our results indicate that the correct document is selected as the first one in only 13% of cases, however, in 40% of cases it is among the top 10 documents and in 75% of cases among the top 100. We believe that this is mainly due to the fact that there is often a significant difference between the description of the same concept (i.e. the text on a Wiki page describing the same concept) across language versions.

Unique relevant links — the results of our system in terms of unique rele-

vant links could be significantly improved by: (a) generating all (1,250) links allowed by the organisers per topic (in most cases our system generated about 200 links per topic). This can be achieved by changing the k parameter of the system controlling the number of articles used as a seed in the target discovery step.

7.1.3 Discussion

The aim in NTCIR-9 CrossLink was to develop a system that is performing well on the provided Wikipedia collection. However, technology for automatic document cross-referencing is also essential in many non-Wiki style document collections. Therefore, it is questionable how easy or difficult it is to apply the developed methods in their context.

The results of the previous link discovery evaluation workshops (Link the Wiki Track: 2007-2010) show that methods relying on the existence of links or semi-structured information that is unique to Wiki-style collections (for example, the correspondence of anchors to article titles) is superior to the methods that are based on purely textual information. Therefore, it is not surprising to see that the majority of the runs submitted to NTCIR 2011: CrossLink were generated by systems exploiting the link and semi-structured information. The organisers of INEX 2009 noticed that algorithms exhibiting high performance on Wikipedia were ineffective on a different Wiki collection (Huang et al.,

2009) (mainly because it was not as extensively linked as Wikipedia and the title information was not as reliable). Similar findings have also been reported by Erbs et al. (2011) who explored link discovery in corporate Wikis and found out that the information about the link graph was not helping the system as much as in the case of Wikipedia. As a result, we believe that link discovery evaluation workshops should in the future more encourage the development of methods that are applicable in a wider context. As these methods are unlikely to perform as well as methods specifically tailored for the collection used in the evaluation, there is currently little incentive to develop them and submit them for evaluation.

At the same time, the development of purely content-based approaches to CLLD constitutes a number of challenges. In particular, (a) these approaches do not allow the use of cross-lingual links between Wikipedia articles — information that has been exploited and found very useful by most of the CrossLink participants, but can hardly be expected to be available in a general context, (b) anchor detection is a hugely challenging problem in a general context as links do not have to be of a conceptual nature (i.e. an anchor is connected to an article which explains it), but can constitute a whole range of semantic relationships (see Chapter 5).

7.1.4 Conclusion

In this section, we have presented and evaluated four methods for Cross-Language Link Discovery (CLLD) applicable to the Wikipedia collection based on the idea of approaching CLLD as a similarity search problem. We have used Explicit Semantic Analysis as a key component in the development of the four presented methods. Our methods produced good results as they ranked in most of the evaluations in which we participated among the top three performers. The results suggest that methods that combine the knowledge of the Wikipedia link graph (including the cross-lingual mapping of articles) with textual semantic similarity can achieve promising results. However, such information is not generally applicable across textual collections and, therefore, it is reasonable to experiment with CLLD methods that operate at the level of textual content.

7.2 KMI @ NTCIR-10 CrossLink-2: CLLD in Wikipedia as a disambiguation and ranking problem

The NTCIR-10 CrossLink-2 task provided an evaluation forum for CLLD methods extending the CrossLink-1 task to more language combinations. This

section discusses the CLLD methods we designed for NTCIR-10 CrossLink-2. The methods approach the CrossLink task as a disambiguation problem making use of some specific properties of Wikipedia. The methods were tested to suggest a set of cross-lingual links from an English Wikipedia article to articles in Chinese, Japanese and Korean (English to CJK) or from an article in Chinese, Japanese and Korean to English (CJK to English). The CJK to English task is new in CrossLink-2. Although tested on these language combinations, the methods are language agnostic and can be easily applied to any other language combination with sufficient corpora and available pre-processing tools. Our methods achieved in the NTCIR-10 CrossLink-2 evaluation the best overall results in the English to Chinese, Japanese and Korean (E2CJK) task and were the top performers in the Chinese, Japanese, Korean to English task (CJK2E)[2] (Tang et al., 2013).

We start by presenting our CLLD methods, explaining the motivation for their design (Section 7.2.1). In Section 7.2.2, we report on the performance of the designed methods. Section 7.2.3 reflects on the findings of this work.

7.2.1 Link discovery methods

Our methods solve the CrossLink task in the steps illustrated in Figure 7.6. Each step is now described in detail.

[2]Our most successful methods in the English to CJK task were not evaluated in the CJK to English task (see Section 7.2.2.1).

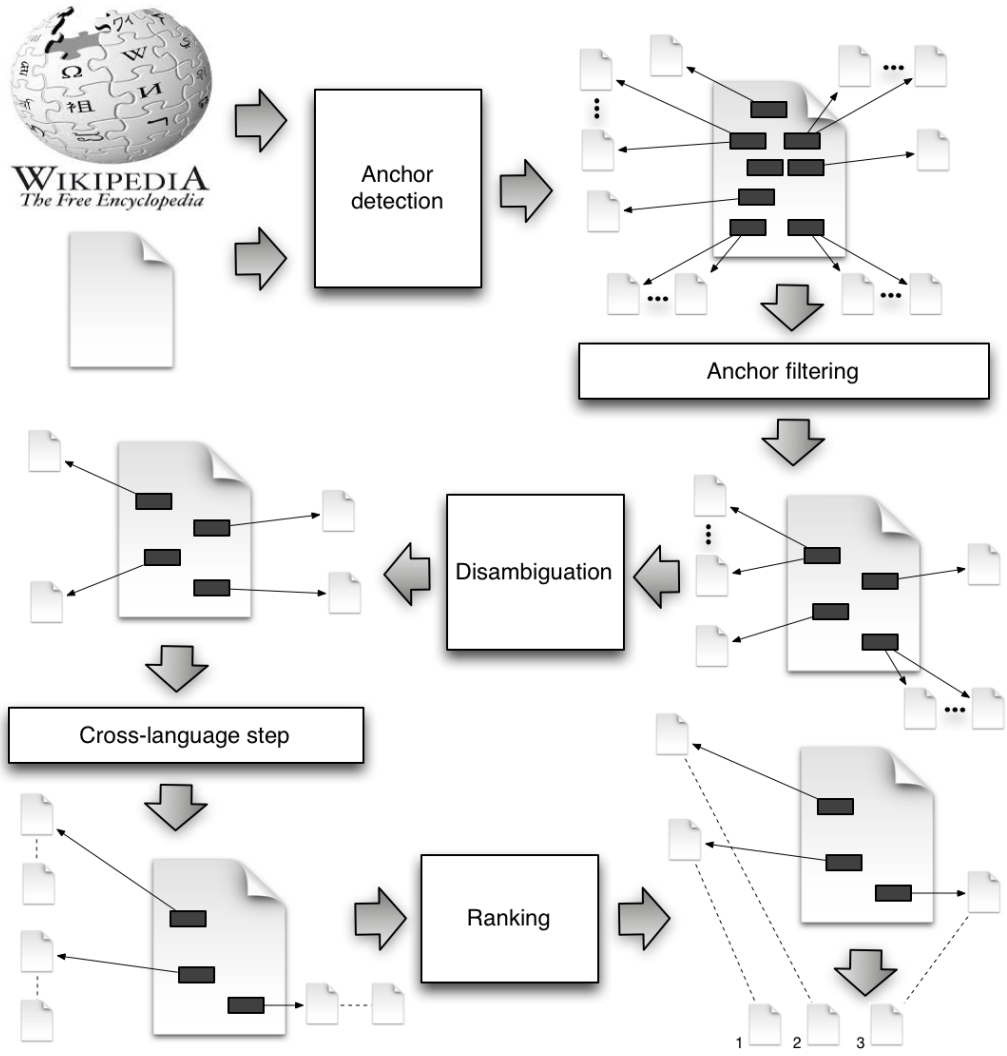


Figure 7.6: The link discovery approach applied in our NTCIR-10 CrossLink-2 methods.

7.2.1.1 Anchor detection

For the purposes of anchor detection, we compiled dictionaries of Wikipedia candidate anchors and concepts for each language. Each anchor corresponds to at least one concept. For example, the English dictionary contains about 14 million terms corresponding to about 4.2 million concepts. We then look up all occurrences of the dictionary terms in the orphan document. To make the anchor detection process quick, we first load the dictionary content into memory using the *trie* data structure and then perform in one pass through the orphan document the identification of dictionary terms in the text of the orphan document.

7.2.1.2 Anchor filtering

The anchor detection step produces many candidate anchors with a very low frequency of occurrence in a general corpus. We measure the prior probability of a term appearing as an anchor to assess how likely a term represents a good anchor. We define this probability as:

$$p(a) = \frac{N_a}{N_t}, \tag{7.1}$$

where N_a is the number of terms t appearing as an anchor a and N_t is the number of terms t in the collection. To make this probability technically easier

to calculate, we estimate it at the granularity of documents. In this case N_a refers to the number of documents where term t appears as an anchor (i.e. term t occurs in the document at least once as an anchor). N_t refers to the number of documents where term t appears. We use an index of the appropriate Wikipedia version to obtain the values of N_a and N_t. Anchors detected in the previous step not satisfying the condition $p(anchor) > \theta$ where θ is a threshold are discarded from further processing. In our runs, we experimentally set this threshold to 0.001.

7.2.1.3 Disambiguation

In the disambiguation step, we select one out of n possible concepts for the detected anchor. The mappings from anchors to concepts is part of the dictionary extracted from Wikipedia we used in the anchor detection step. Using this mapping, given an anchor in one language, we can look up all n possible senses (concepts) of that anchor in that language. This gives us the set of Wikipedia pages the anchor can link to. We calculate a score for each of the available concepts and choose the concept with the highest score.

The scoring measure $s(c, a)$ makes use of two components: (a) the conditional probability of concept c given anchor a and (b) the similarity of anchor's context ctx_a with the text describing concept ctx_c in the source language.

$$s_{c,a} = \alpha p(c|a) + \beta sim(ctx_a, ctx_c), \tag{7.2}$$

where α and β are parameters. While these parameters can be estimated using machine learning techniques to achieve optimal performance, in our runs, we experimentally set $\alpha, \beta = 0.5$ and generally found the system to perform well.

7.2.1.3.1 The probability component

We define the conditional probability of a concept c given an anchor a using the Bayes' rule as follows:

$$p(c|a) = \frac{p(c)p(a|c)}{p(a)}, \tag{7.3}$$

We can estimate $p(a)$ as $p(a) = \frac{N_a}{N_{|A|}}$, where N_a corresponds to the number of occurrences of anchor a and $N_{|A|}$ the number of occurrences of all anchors. We can calculate $p(c)$ as $p(c) = \frac{N_c}{N_{|A|}}$, where N_c is the number of occurrences of (any) anchor representing concept c divided by the total number of occurrences of all anchors $N_{|A|}$. We further estimate $p(a|c)$ as:

$$p(a|c) = \frac{N_{a \cap c}}{N_c}, \tag{7.4}$$

where $N_{a \cap c}$ denotes the number of occurrences anchor a represents concept c. We can then rewrite equation (3) as

$$p(c|a) = \frac{\frac{N_c}{N_{|A|}} \cdot \frac{N_{a \cap c}}{N_c}}{\frac{N_a}{N_{|A|}}} = \frac{N_{a \cap c}}{N_{|A|}} \cdot \frac{N_{|A|}}{N_a} = \frac{N_{a \cap c}}{N_a}. \tag{7.5}$$

7.2.1.3.2 Context similarity component

In our submission, we tested two similarity methods for the purposes of concept disambiguation:

Explicit Semantic Analysis (ESA) – is a method inroduced by Gabrilovich and Markovitch (2007) that calculates semantic relatedness of two texts by mapping their term vectors to a high dimensional space. We have previously explored the use of ESA in the context of link and cross-lingual link discovery in Chapter 6. Since ESA is a method for calculating the similarity of two textual fragments, we apply it to measure the similarity of the context of the anchor being disambiguated with the textual fragments defining the concepts the anchor can be referring to. We define the context of the anchor as the sentence in which the anchor occurs. The context of the concept is defined as the first paragraph of the article describing the concept.[3]

Link Co-occurrence Similarity – calculates the proportion of Wikipedia pages where there occurs both (a) an anchor linking the concept being investigated and (b) an anchor that matches the title of the orphan document. Let t denote the title of the orphan document, c an anchor text referring to the concept investigated and P the set of all Wikipedia pages. We then define the

[3]If the first paragraph is shorter than five sentences we include two or more paragraphs.

link co-occurrence similarity $lcs(t, c, P)$ as

$$lcs(t, c, P) = \frac{|p \in P : t \in p \wedge c \in p|}{|p \in P : t \in p \vee c \in p|}. \tag{7.6}$$

The *lcs* similarity follows the idea that the similarity of two concepts (one representing the orphan document and the second one a Wikipedia page) is expressed by the proportion of Wikipedia pages where both concepts occur together.

7.2.1.4 Cross-language step

The goal of the cross-language step is to find an equivalent concept in the target Wikipedia version to the concept selected in the disambiguation step. In many cases, Wikipedia contains links between pages in different language versions referring to the same concept. In those cases, the cross-language step is straightforward. If a cross-language link is missing for the concept we need to translate, we can make use of the fact that the *same-as* relation is transitive. Therefore, we can try to find the cross-language link using other Wikipedia language versions. For example, there might be no direct cross-language link for translating a concept represented by an English page to Korean, but there might be a link from English to Vietnamese and from Vietnamese to Korean for that concept.

Our implementation uses the following language versions of Wikipedia in

this order: English, German, French, Italian, Dutch, Japanese, Chinese, Korean, Vietnamese. We look for transitive relationships using breadth first search. If a translation for a concept is not found, the concept is discarded from further processing.

While we have observed that having more Wikipedia versions allows us to translate a higher proportion of concepts for the CrossLink language combinations, we believe that a much higher improvement could be seen if the transitivity assumption is applied to language combinations not involving the most resourced Wikipedia language – English.

7.2.1.5 Ranking

In the ranking step, each discovered *(source language) anchor – (target language) concept* pair (link) is assigned a rank. All pairs are then sorted in a descending order according to the their rank and returned in the specified output format (*result list*). Our results show that the ranking phase has a substantial impact on the overall results (see Section 7.2.3). We have experimented with three ranking methods receiving unexpected, but interesting results:

Anchor probability ranking – is a method which assigns as a rank the anchor probability $p(a)$ used in the anchor filtering step (Section 7.2.1.2). Despite its simplicity, this ranking strategy yielded surprisingly good results.

Machine learned ranking – learning optimal ranking from data is a common strategy in information retrieval (Liu, 2009). To test this approach in the context of CLLD, we have extracted a set of features that can be useful for the ranker. We have then trained a ranking Support Vector Machine (SVM-rank (Joachims, 2006)) to learn the optimal ranking model using the pointwise approach. We have then tested all different combinations of the features and also each feature independently. The tested features included:

- *Generality* — the depth of the concept page in the Wikipedia category graph.
- *Category distance* — the shortest path from the orphan document to the concept's page in the category graph normalised by twice the maximum depth.
- *Tfidf* — the term frequency of the term used as an anchor in the orphan document times the inverse document frequency of the concept.
- *Anchor probability* — the anchor probability described in Section 7.2.1.3.1.
- *Similarity* — The ESA or link similarity described in Section 7.2.1.3.2.
- *Relative position* — four features corresponding to the normalised first, last and average position and the position distance of the first and the last occurrence of the anchor in the orphan document.

Surprisingly, we have not seen any combination of these features to outperform in terms of MAP our single best feature - the anchor probability - on its own. Therefore, we decided for simplicity to drop the use of SVM model in our ranking completely. We think this is an interesting negative result. It remains to be determined whether better results can be achieved with these features if the ranking model is trained using the pairwise or listwise approach (Liu, 2009) instead of the pointwise approach.

Oracle ranking – is a non-deterministic approach in which we produce random ranks and test the generated result list against the evaluation tool in the F2F Wikipedia ground truth (GT) setting. The ranking of the best performing result list is then used.

In the experimentation process, we discovered that our methods often generate a low number (significantly less than the allowed 250) but high quality links. Since this can still lead to a decrease in performance, in some of our runs, we top up the result list with additional links until all allowed link slots are used. One strategy is to add alternative disambiguations (i.e. to take the second best, third best, etc. disambiguated concepts for an anchor). We will further discuss this strategy in Section 7.2.3.

7.2.2 Experiments

7.2.2.1 KMI runs

KMI submitted two runs for each E2CJK combination and three runs for each CJK2E combination (15 runs in total). All of the KMI runs follow the pattern described in Figure 7.6 (i.e., 1. Anchor detection, 2. Anchor filtering, 3. Disambiguation 4. Cross-language step, 5. Ranking). The names of the runs code the choices we made in the disambiguation (step 3) and ranking phases (step 5) as described in Table 7.2. The column SIM indicates whether ESA or link similarity (LIS) was used in the disambiguation phase. The column ADD indicates (Y/N) if additional low scoring disambiguations were added in the result set. RANK indicates if oracle (ORC) or anchor probability (APR) were used in ranking. Our CJK2E runs differed from the E2CJK runs in the disambiguation phase. While we used ESA in E2CJK, at the time of submission, we did not have a running instance of ESA for Chinese, Japanese and Korean and therefore used the link similarity approach only.

7.2.2.2 Evaluation

Runs were evaluated in the same way as in CrossLink-1, i.e. using 25 topic documents (different from those used in CrossLink-1). For each topic document there was a maximum limit of 250 anchors that could be extracted,

Run suffix	SIM	ADD	RANK
E2CJK runs			
01-ESA	ESA	Y	APR
02-ORC	ESA	Y	ORC
CJK2E runs			
01-LIS	LIS	Y	APR
02-ORC	LIS	Y	ORC
03-LIS	LIS	N	APR

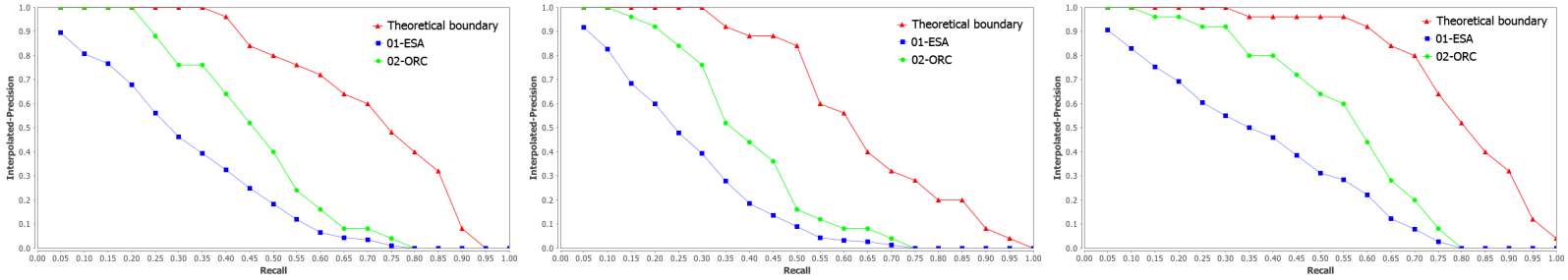
Table 7.2: KMI runs description

each of which could point to the maximum of 5 different target documents (i.e. the maximum number of generated links per topic document was set to 1,250 by the task organisers). The methods have been evaluated at different granularity levels anchor-to-file (A2F) and file-to-file (F2F). There were two evaluation modes: a) GT is derived automatically from the existing link structure of Wikipedia (Wikipedia GT) and b) all anchors and targets are pooled and the evaluation is carried out by a human assessor (Manual assessment). *Precision-at-N (P@N)*, *R-Prec*, and *Mean Average Precision (MAP)* were used as the main performance metrics. More information about GT, the evaluation setup and a detailed description of the evaluation measures can be found in the overview paper (Tang et al., 2013) and Section 7.3.

The results for all experiments, including a theoretical boundary for F2F Wiki GT explained in Section 7.3, are summarised in Table 7.3. Graphs 7.7, 7.8, 7.9, 7.10, 7.11, 7.12 show the performance of the designed methods for different language combinations and assessment strategies.

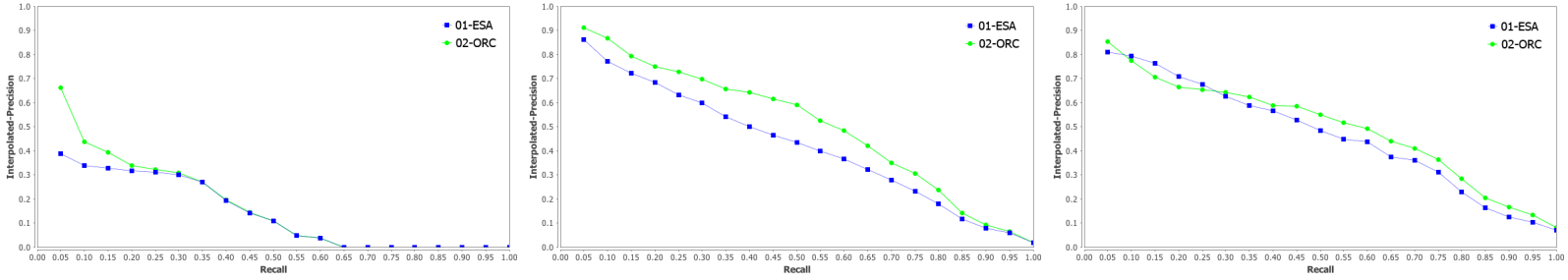
Run ID	LMAP	R-Prec	Run ID	LMAP	R-Prec
English-to-Chinese			**Chinese-to-English**		
F2F, Wikipedia ground truth			F2F, Wikipedia ground truth		
KMI-E2C-A2F-02-ORC	0.404	0.404	KMI-C2E-A2F-02-ORC	0.221	0.337
KMI-E2C-A2F-01-ESA	0.249	0.335	KMI-C2E-A2F-01-LIS	0.221	0.336
			KMI-C2E-A2F-03-LIS	0.219	0.336
F2F, manual assessment			F2F, manual assessment		
KMI-E2C-A2F-02-ORC	0.133	0.273	KMI-C2E-A2F-02-ORC	0.067	0.180
KMI-E2C-A2F-01-ESA	0.112	0.275	KMI-C2E-A2F-01-LIS	0.067	0.180
			KMI-C2E-A2F-03-LIS	0.064	0.180
A2F, manual assessment			A2F, manual assessment		
KMI-E2C-A2F-01-ESA	0.174	0.201	KMI-C2E-A2F-01-LIS	0.077	0.060
KMI-E2C-A2F-02-ORC	0.168	0.210	KMI-C2E-A2F-02-ORC	0.077	0.060
			KMI-C2E-A2F-03-LIS	0.076	0.060
English-to-Japanese			**Japanese-to-English**		
F2F, Wikipedia ground truth			F2F, Wikipedia ground truth		
KMI-E2J-A2F-02-ORC	0.341	0.341	KMI-J2E-A2F-02-ORC	0.224	0.224
KMI-E2J-A2F-01-ESA	0.206	0.285	KMI-J2E-A2F-01-LIS	0.114	0.176
			KMI-J2E-A2F-03-LIS	0.113	0.176
F2F, manual assessment			F2F, manual assessment		
KMI-E2J-A2F-02-ORC	0.450	0.513	KMI-J2E-A2F-02-ORC	0.171	0.271
KMI-E2J-A2F-01-ESA	0.383	0.424	KMI-J2E-A2F-01-LIS	0.138	0.202
			KMI-J2E-A2F-03-LIS	0.137	0.202
A2F, manual assessment			A2F, manual assessment		
KMI-E2J-A2F-02-ORC	0.452	0.337	KMI-J2E-A2F-02-ORC	0.072	0.058
KMI-E2J-A2F-01-ESA	0.440	0.279	KMI-J2E-A2F-03-LIS	0.062	0.042
			KMI-J2E-A2F-01-LIS	0.062	0.042
English-to-Korean			**Korean-to-English**		
F2F, Wikipedia ground truth			F2F, Wikipedia ground truth		
KMI-E2K-A2F-02-ORC	0.492	0.492	KMI-K2E-A2F-01-ORC	0.144	0.240
KMI-E2K-A2F-01-ESA	0.302	0.384	KMI-K2E-A2F-03-LIS	0.143	0.240
			KMI-K2E-A2F-01-LIS	0.143	0.239
F2F, manual assessment			F2F, manual assessment		
KMI-E2K-A2F-02-ORC	0.433	0.493	KMI-K2E-A2F-01-ORC	0.264	0.284
KMI-E2K-A2F-01-ESA	0.424	0.457	KMI-K2E-A2F-01-LIS	0.262	0.284
			KMI-K2E-A2F-03-LIS	0.260	0.284
A2F, manual assessment			A2F, manual assessment		
KMI-E2K-A2F-01-ESA	0.537	0.311	KMI-K2E-A2F-01-LIS	0.184	0.073
KMI-E2K-A2F-02-ORC	0.533	0.293	KMI-K2E-A2F-01-ORC	0.184	0.073
			KMI-K2E-A2F-03-LIS	0.180	0.073

Table 7.3: The summary of the KMI runs results.



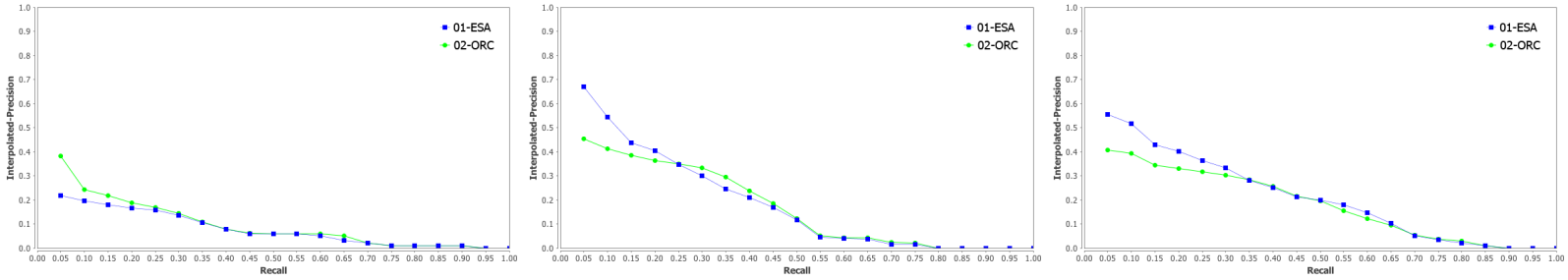
(a) Chinese (b) Japanese (c) Korean

Figure 7.7: E2CJK F2F evaluation results with Wikipedia ground truth



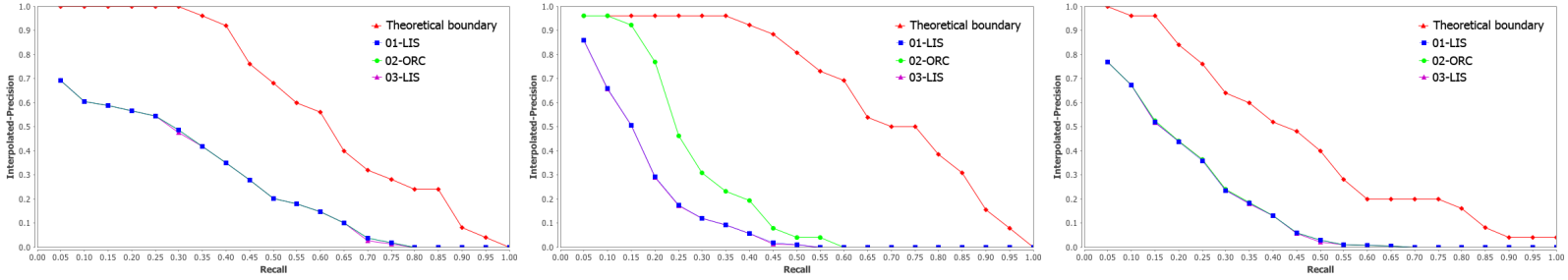
(a) Chinese (b) Japanese (c) Korean

Figure 7.8: E2CJK F2F evaluation results with manual assessment



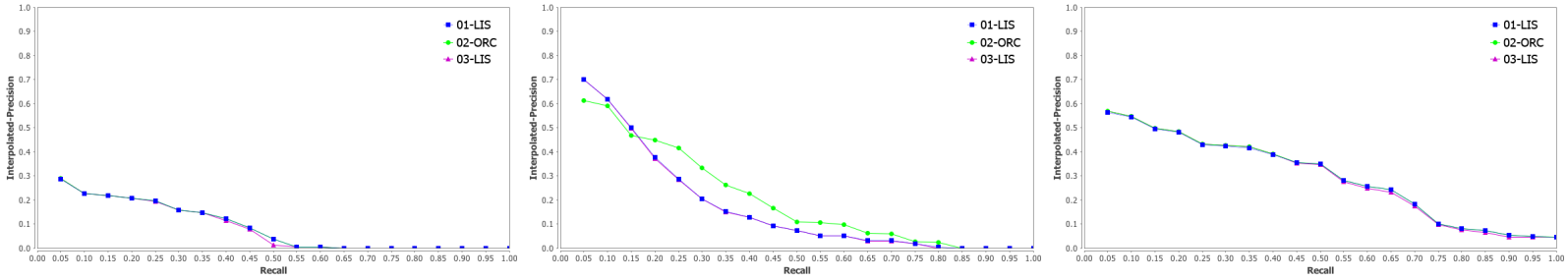
(a) Chinese (b) Japanese (c) Korean

Figure 7.9: E2CJK A2F evaluation results with manual assessment



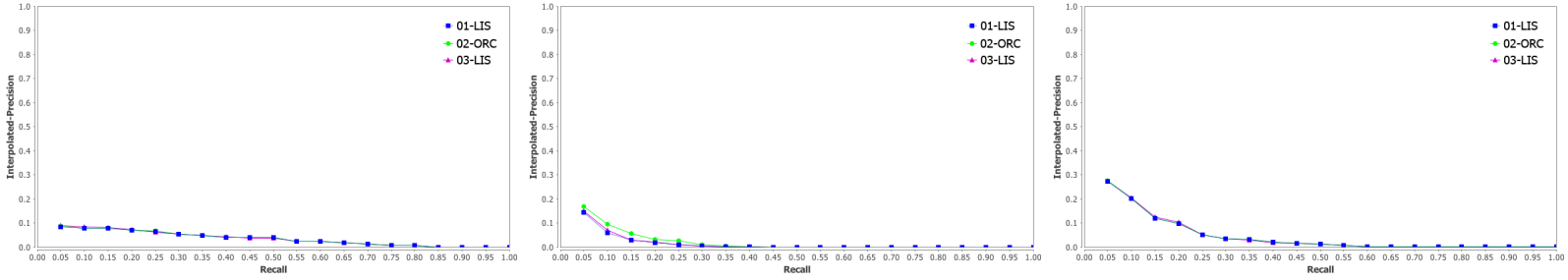
(a) Chinese (b) Japanese (c) Korean

Figure 7.10: CJK2E F2F evaluation results with Wikipedia ground truth



(a) Chinese (b) Japanese (c) Korean

Figure 7.11: CJK2E F2F evaluation results with manual assessment



(a) Chinese (b) Japanese (c) Korean

Figure 7.12: CJK2E A2F evaluation results with manual assessment

7.2.2.3 Performance comparison with other teams

The NTCIR-10 CrossLink-2 organisers reported (Tang et al., 2013) that overall, our methods achieved the highest scores in multiple evaluation scenarios (measured with different metrics: *LMAP, R-Prec, Precision-at-N* in different evaluation levels against different GTs) for E2CJK. KMI methods are consistently the top (mostly among the top three) performers in the CJK2E task. In total, 67 runs from 10 teams were evaluated.

7.2.2.4 How can the performance be improved?

There is a number of ways in which our methods could be improved and optimised for better performance. We see the main possibilities in:

The use of ESA for disambiguation in CJK2E – Our methods utilised ESA only in E2CJK tasks where it performed consistently better than link similarity, which was used in all CJK2E experiments. Yet, ESA can be in a straightforward way adapted for Asian languages.

Anchor detection - We have compared our results with the runs of other teams and discovered that our system did not detect anchors that were only part of a term and we also did not use stemming. For example we did not detect anchor *plaque* in term *plaque-reducing* and anchor *Korea* in term *Korean king* (while links to *Dental plaque* and *Korea* were in GT). In English Wikipedia, anchors that are not composed of whole tokens do not exist, but it remains

to be determined whether generating them can be useful. In addition, we discovered that in the distributed orphan documents end of line characters were often missing, which resulted in the concatenation of some words, such as *poultry* into *poultryand*, and also not all markup was removed, which is why we did not detect anchor *Peking duck* in string *Peking duck.JPG*. Consequently, the fact that our anchor detection algorithm assumed anchors to be composed only of whole terms had a significant negative impact on the performance of our runs.

Tuning parameters in the disambiguation step – In our submission, we have set the parameters α and β used as weights for the similarity and probability components in the disambiguation stage as equal, however we expect it would be possible to tune (or machine learn) these parameters to achieve better performance. Such approach would be similar to the one reported in (Milne and Witten, 2008).

Considering more than one disambiguation per anchor in the first step – There are many situations when it makes sense for an anchor to link to multiple targets. For example, in the context of an article about American War, it can be relevant for the anchor *president* to link to the page explaining the general concept, the page about the *President of the United States of America* as well as the page about the 16th president of the United States *Abraham Lincoln*. While the Web (HTML) does not support by default multiple links per anchor,

such approach can be easily put into practise and has been encouraged by the task organisers. Our implementation of the methods currently selects the best disambiguation in the first round and the second best, third best, etc. in the following rounds after the best, second best, etc. disambiguated concept is selected for each anchor. It might be possible to achieve better performance in manual assessment if more than one disambiguation is used in the first round. However, it is likely this would decrease the performance of the system in the Wiki evaluation as there is by definition only a single link per anchor in Wiki GT.

7.2.3 Discussion

There are two main outcomes that follow from the evaluation results:

ESA vs link similarity disambiguation – Our experiments show that ESA outperforms link similarity.

Ranking strategy – Ranking is a CLLD subtasks which has perhaps the highest influence on the final results. It is interesting that in our case, a trivial ranking technique produced better results than the SVM machine learned model using the pointwise approach. Regardless of whether better ranking features can be found and whether a better model can be trained using the pointwise or listwise approach, we believe that in order to develop more optimal ranking strategies, it is crucial to better understand the nature of the methods

(where does the system make mistakes) and the task itself (what is exactly in GT). Our results demonstrate that while the optimal ranking techniques (ORC runs) with one GT (for which they were optimised) achieve substantially higher performance than our anchor probability ranking runs, the ESA runs perform equally well when applied to a different GT. This suggests the following: (a) the quite simple anchor probability ranking is almost as good as the oracle ranking leaving little room for improvement of ranking methods unless we want to over-fit them to achieve high performance on one particular GT, (b) it confirms how subjective the CLLD task is (Chapter 6) and largely explains the fairly high variability of the results of different systems under different evaluation settings.

7.2.4 Conclusion

In this section, we presented CLLD methods we submitted for evaluation at NTCIR-10 CrossLink-2. The methods of our team achieved the best results in the E2CJK task and were the top performer in the CJK2E task, where we did not make use of such a solid disambiguation system as we deployed in the E2CJK task. The experiments carried out indicate, primarily, that ESA outperforms link similarity in the disambiguation phase and show the high importance of ranking on the overall results, with some interesting unexpected results on well performing ranking methods. Comparing the overall results we

achieved at NTCIR-10 with those at NTCIR-9, we can see a very significant increase in MAP. We believe this is mainly due to the reliance of our NTCIR-10 algorithms on the disambiguation information in Wikipedia, which we did not use in NTCIR-9 and which substantially narrows down the space of possible link targets.

7.3 Discussing the evaluation methodology

Choosing the right evaluation methodology is certainly one of the greatest challenges in link discovery. Suitable interlinked corpora that could be used for evaluation are lacking and creating it manually would require huge effort. It has been previously reported by Huang, Xu, Trotman and Geva (2008) that Wikipedia should not be seen as a reliable ground truth. When establishing the ground truth based on the link structure of different language versions, we can see that the correlation of their link structures is surprisingly low (Chapter 6). Therefore, the automatic assessment results should be treated as informative only. At the same time, care should also be taken when interpreting the manual assessment results as the task of interlinking content can be considered highly subjective (see Section 3.3.2).

The existence of a good evaluation framework which makes it possible to recognise and justify (both major and minor) improvements to the methods or reject method updates that do not improve performance is critical to the

continuous technology progress of link discovery systems. A good evaluation framework will have the characteristic of assigning a system that produces better results a higher score than to a systems that produces worse results. This behaviour will be primarily stable (consistent from one set of topics to another) and reliable (an improvement in score will truly correspond to an improvement in user experience). The key to designing such an evaluation framework is to understand what is expected from an ideal system. The resemblance of the system's characteristics to the characteristics of the ideal system should then be captured by the framework as accurately as possible.

Since the system output is in CrossLink defined as a ranked list of anchor-target pairs, the performance of two systems can be compared by assessing their ranked lists. To do this, an evaluation framework will typically define (a) the set of (possibly graded) correct answers (GT) and (b) the methods for calculating the score based on the system's answers (evaluation metrics). The CrossLink evaluation task (Tang et al., 2013) defines two GTs, the automatic (Wiki) and the manual assessment, and a set of evaluation measures, which are based on standard information retrieval metrics (MAP, R-Prec, Precision-at-n) and are applied on the participants' result sets at the A2F or F2F granularity.

Some of the limitations of the current evaluation approach, such as the inaccuracy/subjectivity of the Wiki GT, have already been widely known by both the participants and the task organisers. However, as we were designing

and evaluating our methods for CrossLink-2, we identified a few more evaluation pitfalls about which we informed the task organisers. From our email conversation, it became clear that even they did not have a unanimous view on how these issues should be approached. As the knowledge of these issues contributes to the better understanding of the link discovery task, we discuss them here and propose how the evaluation framework can be improved in the future.

7.3.1 GT definition

The Wiki GT set for a given Wiki page (*topic*) is defined in CrossLink as the union of the concepts linked from either the source or the destination Wiki version (the source language concepts are mapped to their equivalent concepts in the destination Wiki version). Since equivalent pages in different Wiki versions often provide substantially different information on the same topic, there is consequently a low correlation (typically less than 0.2) of their respective link structures (Chapter 6).

Therefore, the current approach has certain disadvantages one should be aware of. 1) An ideal system that will correctly identify all relevant anchors in the orphan document and will correctly link them to their relevant concepts in the destination Wikipedia version will not achieve 100% recall, because there is typically a large set of links in GT for which no relevant anchor in the or-

phan document exists. 2) Since Wiki GT evaluation is carried out only at the F2F level, a possible way how to achieve close to 100% recall would be to guess concepts, which are linked in the target language version of the orphan document and for which there does not exist any relevant anchor in the source document, and assign them any (even irrelevant) anchor in the orphan document. Although this strategy could potentially lead to better performance, we think it should be discouraged as it exploits a particular weakness in the evaluation methodology and changes the meaning of the CrossLink task.

7.3.2 The theoretical performance boundary

The findings reported in the previous section led us to measure the theoretical boundary in CrossLink-2 (F2F evaluation with Wiki GT). This boundary gives us the performance of an ideal system, which is constructed as follows: we take the original GT and remove from it all target language concepts for which there does not exist any relevant term (or even substring of a term) in the orphan document that could be used as an anchor pointing to this concept. The run submission is then constructed only from the remaining (correct) concepts in GT. The idea of the theoretical boundary is to find the maximum performance a CLLD system can achieve in this task. The calculation of the theoretical boundary is based on the November 2012 dump of Wikipedia with the CrossLink-2 GT. Although the calculated theoretical boundary can slightly

change according to the Wikipedia version used, we consider the produced boundary depicted in Figures 7.7 and 7.10 sufficiently accurate for the purposes of the CrossLink-2 evaluation. We believe that comparing the submitted runs with the theoretical boundary is more informative of systems' performance than the absolute evaluation scores. While the achieved absolute scores might seem in many cases quite low, it is possible to see from the comparison that, in particular in the E2CJK task, the performance of the CLLD systems is actually fairly good.

7.3.3 Ranking largely determines performance

We experimented with different ranking strategies for Wiki GT including the extreme cases where a system gets all the correct answers on the top or the bottom positions in the result list. We observed that ranking largely influences how successful a system is in the evaluation. Typically, by changing the order of anchors in the output file, we were able to get LMAP corresponding to both a top performing system as well as a system at the bottom of the evaluation chart. It directly follows from the way how LMAP is calculated that providing correct answers on the top positions is critical. Consequently, one of the problems with the application of LMAP in CrossLink is that the GT is unstable/subjective and the retrieved links are not equal, because some of the links are much more relevant than others. For example, in an article

about *Japan* the link to *Tokyo* is certainly more important than the link to the *Michelin Guide*, yet systems are rewarded in the same way for retrieving any of them. This can lead to situations where systems with very different qualitative properties are assigned the same LMAP score. We think that a way to mitigate this issue (apart from the already used manual assessment) would be to apply one of the existing graded relevance evaluation metrics (Sakai, 2009). The graded GT could be constructed as a multiset union of links in all Wikipedia languages (instead of a set union of the two considered languages). We think this approach would not only lead to more informative results, but might also help stabilise the fluctuations in results of participants in different language combinations and evaluation settings.

7.3.4 The evaluation metric rewards certainty, not relevance

CrossLink aims to encourage the development of systems that can link an anchor to multiple concepts. The reasons why this is useful are explained in Section 7.2.2.4. Consequently, the run submission format allows participants to report more than one target concept per anchor. However, the only allowed way of expressing this is by assigning all the concepts associated with that anchor a single position in the result list. This means, for example, that a system can in an article about *India* generate anchor *Gandhi* with links to

Mahatma Gandhi, *Gandhi (film)* and *Gandhi (American band)* and must assign them a single position in the result list. The first link is certainly correct, the second link seems useful and the third link is certainly incorrect. The problem of this approach is that: A system (a) cannot provide any ranking for the generated concepts, i.e. all concepts are treated equal and the correctness of the anchor is evaluated according to Equation 5 in (Tang et al., 2013) as the proportion of those concepts that were correct and (b) cannot decide to link a concept with high relevance for a given anchor, then generate other anchors and eventually additional concepts with lower relevance for the given anchor.

Since the performance of a system is critically influenced by the links generated in the first positions, this leads to a situation in which systems are encouraged to first generate "low risk" anchors. Unambiguous anchors, which by their nature are difficult to get wrong, constitute this low risk. Therefore, an effective strategy is to choose less relevant, but certain anchors, before highly relevant but ambiguous anchors. As acknowledged by one of the organisers, the problem is that this approach rewards certainty, not precision. Also, according to Equation 5 in (Tang et al., 2013), a system cannot be rewarded for generating more than one target per anchor as from a strategic point of view, it is better to select one concept (about which a system is the most certain) rather than more concepts. The solution would be to allow the ranking in the output file at the granularity of targets (rather than at the granularity

of anchors).

7.4 Summary of contribution

The main goal of this chapter was to design and evaluate new link discovery methods in the context of an international evaluation conference (Goal 1). We also aimed to address the noun phrase-to-document CLLD problem (RQ 3) and to discuss the ways to evaluate and interpret the performance achieved by link discovery methods (RQ 4).

We participated twice in the CrossLink evaluations, namely at NTCIR-9 CrossLink and NTCIR-10 CrossLink-2, submitting and evaluating novel CLLD methods and achieving very good results. Our methods were among the top performers in NTCIR-9, having the highest early precision under manual assessment (A2F P@5) (Tang et al., 2014) and being in most categories among the top three systems. In the NTCIR-10 CrossLink-2, we have achieved the overall best results in the English to CJK categories (Tang et al., 2013) and were the top (steadily among the three best and mostly second best) performer in the CJK to English task.

The fact that we used two different strategies in each evaluation allows us to draw some interesting conclusions based on our experience and the evaluation results. In NTCIR-9 CrossLink, our methods were based on the idea of similarity search. As reported by the task organisers, these methods proved

to be effective in suggesting novel links that allow the link graph to grow, as opposed to, for example, the methods of QUT (Queensland University of Technology) (Tang, Cavanagh, Trotman, Geva, Xu and Sitbon, 2011) that link mine primarily for ensuring consistency across the collection (Tang, Itakura, Geva, Trotman and Xu, 2011). We were also the only to test cross-language similarity search in one of our runs, instead of relying on the explicit cross-language links available in Wikipedia. This indicated a substantial difference in performance when the explicit information is not available. As this is the case in most real-world collections, this suggests that CLLD evaluation conferences should in the future consider whether to provide a separate evaluation for methods that do not depend on this information. There are two main contributions of our NTCIR-9 CrossLink methods. Firstly, while our approach has been significantly different from that of other evaluation participants, the methods have demonstrated good performance and the ability to discover novel links. Secondly, we have shown the methods have the potential for being applied in different multilingual collections beyond Wikipedia.

In NTCIR-10 CrossLink-2, we have applied a different strategy. We treated the task as a disambiguation problem and focused on the ranking subtask. While we were not surprised to see that this strategy yielded better results, as exploiting the information about term senses significantly narrows down the space of possible link targets, we have provided a proof of concept of the

efficiency of this approach in Wikipedia. The main surprise of this work, however, was to see a fairly simple ranking method based on anchor probability outperform a more sophisticated ranking method based on machine learning. While finding the reasons for this is out of scope of this work, we speculate this might potentially be an artefact of the evaluation framework preferring strategies ranking based on confidence rather than relevance. The main contribution of our NTCIR-10 CrossLink-2 methods is in both the design of a novel method, which uses ESA in the disambiguation step, and in showing that an application of a trivial anchor probability ranking method can yield very good results. This is, in particular, interesting as we have shown the substantial impact of the ranking phase on the overall results, which is something we were not previously aware of.

Apart from the methods design, an important part of our contribution is in the area of the evaluation framework. As the problems of evaluation against the Wikipedia ground truth, which we already discussed in Chapter 6, are also valid in this context, we further analysed this problem. Motivated by the issue of the inability of an ideal system to achieve perfect precision and recall due to the comparative but not parallel nature of Wikipedia translations, we have decided to measure the theoretical performance boundary. We think this is a particularly important contribution as we believe that the ability to compare the results with the theoretical boundary is more informative of a

system's performance than the absolute evaluation scores. Finally, we have also suggested the use of graded relevance metrics in future evaluations, as this would motivate the development of methods that rank based on relevance rather than confidence, and highlighted the issues in the current output format, which do not motivate the submission of runs ranking according to relevance.

In the next chapter, we will come back to the problem of monolingual link discovery. We will focus on the problem of facilitating access in a large distributed document collection of research papers, which does not have as many semi-structured features as Wikipedia. This work enables to experiment with large scale link discovery in a dataset where these techniques can provide a wide range of benefits.

Chapter 8

From Link Discovery to Knowledge Discovery: Towards an Improved Access to Public Knowledge Through Aggregations and Link Discovery

The previous chapters dealt with various link discovery methods, their adaption to multilingual and domain-specific settings, and issues related to their evaluation. This chapter has in this sense very different goals.

At the beginning of the thesis (in Chapters 1 and 2), we discussed the limitations of accessing knowledge inherent in large document collections using traditional search approaches, which are typically based on submitting a

keyword query to a search engine. We explained that there is a need for supporting exploratory access to these large collections in order to help uncover (unknown) relationships. We argued that knowledge is expressed in associations/relationships connecting pieces of information. As shown by Swanson (1986) (see Section 3.2.1.3), these relationships might be either explicitly stated in documents (they are already known) or remain hidden (undiscovered), typically as a result of the different pieces of the puzzle being spread across documents.

As a consequence, one of the most (if not the most) important application of link discovery methods is in improving the access to knowledge contained in research papers. Research papers are the medium for communicating the results of research in all disciplines. The work in this chapter builds on the premise that the application of link discovery techniques to automatically connect semantically related research work within and across disciplines will improve access to human knowledge, resulting in substantial benefits to society.

However, in order to apply link discovery to this collection, it is first necessary to have access to the data from many systems, which can be achieved by aggregating these data. Such aggregation did not exist at the time this research work was undertaken (see Section 8.2) and its development was therefore necessary. Consequently, we will discuss the challenges in machine access to these data, as they are making it more complicated to exploit the benefits of link

discovery in practice. Since carrying out this work faces the limitations implied by copyright law, our work focuses on building an aggregation of Open Access (OA) articles only, i.e. freely available research publications on the Internet with both access and reuse rights.

As the development of the aggregation benefits users beyond the link discovery community, this chapter will deal with the wider context of developing an aggregation of OA research papers, yet link discovery will receive some special attention. We will present link discovery as a component of a larger system which fulfills a number of more general use cases, but where link discovery contributes to its overall mission. In particular, we will explore a use case in which link discovery is applied in a digital library system to provide browsing facilities on an originally distributed collection of millions of documents. The system has been developed as part of this research work, as a necessary prerequisite, to be able to demonstrate the value of link discovery in an important application domain, show link discovery methods are scalable to very large textual collections and to discuss ways in which the generated links can be presented to the user.

The chapter addresses Goal 2: *Show that link discovery techniques can be deployed in large document collections to facilitate access to public knowledge.*

8.1 The opportunity to exploit Open Access content

The last 10 years have seen a massive increase in the amount of OA publications in journals and institutional repositories, indicating a culture change happening across research disciplines. This culture change has been supported at different levels by governments and funders, digital libraries, universities and research institutions, and last, but not least, by researchers and the general public. The number of OA journals and institutional repositories is increasing year by year and the number of OA papers grows exponentially. The OA movement is certainly gaining momentum.

Apart from the growth of OA journals, providing so-called *Gold OA*, a cornerstone of the OA movement has been a process called self-archiving (*Green OA*). Self-archiving refers to the deposit of a preprint or postprint of a journal or conference article in an institutional repository or archive. According to the Directory of Open Access Repositories (OpenDOAR)[1], more than 65% of publishers endorse self-archiving. They include major players, such as Springer Verlag. It is largely due to this policy that millions of research papers are available online without the necessity to pay a subscription.

The presence of this freely accessible and quickly growing knowledge pool

[1] http://www.opendoar.org/

constitutes a great opportunity as well as a challenge. We believe that the availability of open data creates an environment in which technology innovation can accelerate. Open availability of data has been the defining feature of the Web which largely contributed to the success of search engines, wikis, social networks and many other online services. OA appears to be going in the same footsteps. If the current environment is supportive of Open Access, is there anything else that reduces its impact and makes it difficult for OA to become the default policy? There is still, of course, a number of issues that hinder the adoption of OA (Björk, 2003) including often discussed legal barriers as well as the issues related to evidence of scientific recognition. In this chapter, we discuss a very important yet a rarely debated one - the lack of a mature technical infrastructure.

The Budapest Open Access Initiative[2] clearly identifies, in its original definition of Open Access from 2001, that OA is not only about making research outputs freely available for downloading and reading. The aspect of reuse, which includes being able to index and pass OA content to software, is firmly embedded in the definition, opening new possibilities for the development of innovative OA services, such as those based on Text and Data Mining (TDM). However, while the growth of OA content has been used in the last decade as a benchmark of success of the OA movement, the successes in terms of finding and reusing OA content are much less documented. We believe that

[2] http://www.budapestopenaccessinitiative.org/

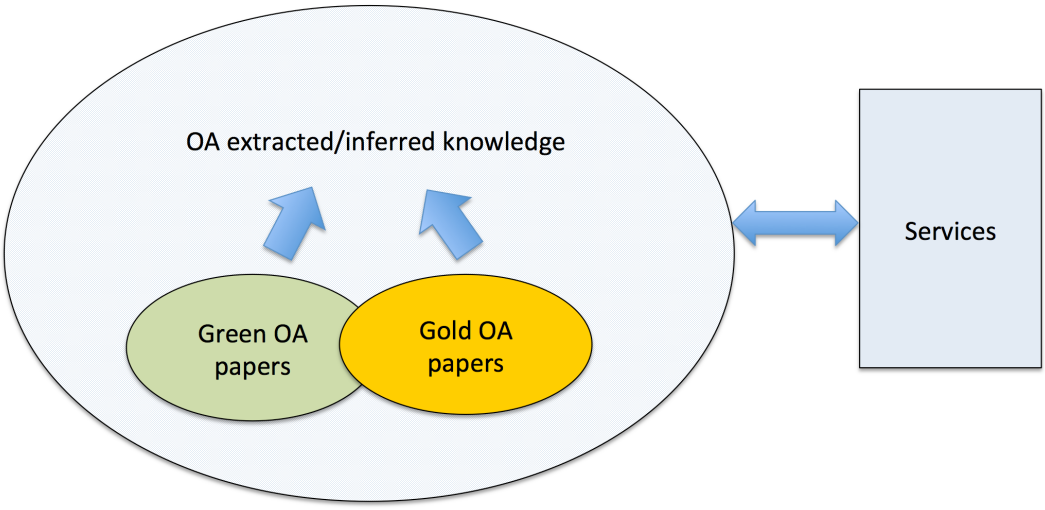


Figure 8.1: Open Access content and services.

in order to fully exploit the reuse potential of OA, it is vital to improve the current OA technical infrastructure and facilitate the creation of novel (possibly unforeseen) services utilising the OA content. According to Figure 8.1 below, this OA content consists of (both Green and Gold) OA papers and all possibly inferred or extracted knowledge that is expressed in these materials. The services that can access and manipulate this content can be tailored to different audiences and serve different purposes.

There are three essential types of data access, which we will call *access levels*. We argue that these access levels must be supported by services in order to create an environment in which the potential of OA content can be fully exploited. They are:

- Programmable machine access to raw data[3].

[3]The concept of raw data refers in this context to structured or unstructured publication

- Access at the granularity of papers.

- Analytical access at the granularity of collections.

In this chapter, we introduce them and demonstrate why their combination is vital. Later, we present the CORE (COnnecting REpositories) system which delivers services at these access levels and provides the infrastructure to facilitate the creation of new services, including those utilising TDM, such as link discovery.

The rest of the chapter is organised as follows. Section 8.2 reviews the related work in the area of research paper aggregations. In Section 8.3, we analyse the functionalities an OA technical infrastructure should provide. We then introduce the layered model of an aggregation system and use it to highlight the functionality currently not provided by existing aggregation systems. Section 8.4, introduces the CORE system and discusses the technical issues in delivering an infrastructure that embraces the above mentioned principles. We also describe how link discovery has been integrated into CORE and explain our work in using the discovered document-to-document links to provide exploratory search. Finally, we outline the future work and provide a discussion of the implications.

---

manuscript data provided by repositories or resulting from processing at the level of aggregation to be used for further machine processing. Our concept of raw data is different from the currently often discussed concept of *research data* which typically refers to data in the form of facts, observations, images, computer program results, recordings, measurements or experiences in which an argument, test or hypothesis, or another research output is based.

8.2 The need for an Open Access aggregation

The recent years have seen a substantial investment in the development of Open Access Repositories (OARs) supporting the deposit and preservation of Open Access content. While the existence of these repositories is crucial, they are, however, only one side of the coin. The Confederation of Open Access Repositories states:

> "Each individual repository is of limited value for research: the real power of Open Access lies in the possibility of connecting and tying together repositories, which is why we need interoperability. In order to create a seamless layer of content through connected repositories from around the world, Open Access relies on interoperability, the ability for systems to communicate with each other and pass information back and forth in a usable format. Interoperability allows us to exploit today's computational power so that we can aggregate, data mine, create new tools and services, and generate new knowledge from repository content" – (Rodrigues and Clobridge, 2011).

While we fully acknowledge the importance of interoperability, it is in fact the implementation of interoperability in the form of systems and services that will enable us to derive knowledge from the information stored in the

form of articles, books, thesis, etc., in OARs. The technical maturity of the infrastructure for connecting and tying together repositories is vital for the success of OA.

A number of projects/systems have addressed the issue of connecting OARs by developing metadata repository aggregators, such as BASE (Pieper and Summann, 2006), IRS (Lyte et al., 2009) or OAISter (Loesch, 2010). The majority of repository aggregation systems focus on the essential problem of aggregating resources for the purposes of providing cross-repository metadata search. While search is an essential component of an OA infrastructure, connecting and tying OA repositories together offers far more possibilities. Aggregations should not become just large searchable metadata silos, they should offer (or enable others to offer) a wide range of value-added services targeting all types of users participating in the research process. That is not just users looking-up individual publications to read, but, for example, those who need to explore certain research areas, access statistical information about collections of publications and trends as well as those who need machine access to publications to carry out experiments and develop new services. These characteristics distinguish OA aggregation systems from major academic search engines, such as Google Scholar or Microsoft Academic Search.

More specifically, Table 8.1 shows the support provided by Google Scholar and MS Academic Search at the three above mentioned access levels. As

Google Scholar does not provide an API at all and the API provided by MS Academic Search is restricted in the way it can be used, these systems offer only very limited support for those wanting to build new systems on top of them or for those who need machine access to the data. Due to the lack of data analytic use cases they support and the restrictions on machine access to the data, both services are difficult to apply for carrying our various forms of data analysis without breaching their Terms & Conditions.

Access level	**Google Scholar**	**MS Academic Search**
Programmable access to raw data	No API is provided, scraping solutions exist, but are likely to be illegal.	Non-commercial use only, max 200 queries per minute, only first 100 items can be accessed. Not to be used to crawl the corpus.
Access at the granularity of papers	The Google Scholar search interface.	The MS Academic Search interface
Analytical access at the granularity of collections	No specific services	No specific services

Table 8.1: Access levels, as defined in Section 8.1, provided by the two major commercial academic search engines.

The idea of going "beyond search and access" while not ignoring these functions has already been explored by Lagoze et al. (2005). The authors argue that digital libraries need to add value to web resources by extending current metadata models to allow establishing and representing the context of resources, enriching them with new information and relationships and encourage collaborative use. While the value of such services is apparent, and their realisation is often entirely plausible, there is a high barrier in entering the market. This barrier is in the difficulty of being able to access and work with the data needed to realise these services.

As highlighted by the experience of the OAIster team, the realisation of traditional solutions to aggregation systems tends to be expensive and the resulting production systems are hard to sustain in the long term (Manghi, Mikulicic, Candela, Castelli and Pagano, 2010). Therefore, aggregation systems must either (a) become significantly cheaper to develop and run or (b) there should be an open infrastructure that allows others to build on top of the aggregated content.

Option (a) has been recently addressed by the team developing the D-NET architecture (Manghi, Mikulicic, Candela, Artini and Bardi, 2010). The D-NET software is realised as a service-oriented architecture providing services ranging from metadata harvesting and harmonisation to the way resources are presented. However, even with reusable software packages, significantly

reducing the cost of aggregation is not a trivial task given the growing amounts of content available online and the need to store aggregated data. Therefore, Option (b) focuses on offering an open yet a ready-to-use solution in the form of an open online service. This principle is embraced by the CORE system presented in this chapter.

One of the important aspects which D-NET shares with CORE is the aggregation of both content and metadata as opposed to the previously mentioned metadata only aggregations. CORE makes already substantial use of the full-text content for various purposes including citation extraction and resolution, link discovery (content recommendation and deduplication), content classification and others. This allows us to build services that clearly add value to the content provided by individual repositories. The D-NET framework is in this sense going in the same direction as CORE by promising to implement these services in the future.

8.3 The users and layers of aggregations

According to the level of abstraction at which a user communicates with an aggregation system, it is possible to identify the following levels of access:

1. *Programmable (raw) data access*

2. *Transaction access*

3. *Analytical access*

With these access levels in mind, we can think of the different kinds of users of aggregation systems and map them according to their typical access level. Table 8.2 lists the kinds of users that we have identified as the main players in the OA ecosystem and explains how aggregations can serve them. Naturally, each user group will expect to communicate with an aggregation system in a specific way that will be most practical for satisfying their information needs. While developers are mostly interested in accessing the data, for example through an API or a data dump, individuals will primarily require access to the content at the level of individual items or relatively small sets of items, mostly expecting to communicate with a digital library (DL) using a set of search and exploration tools. A relatively specific group of users are eResearchers[4] whose work is largely motivated by information communicated at the transaction and analytical levels, but in terms of their actual work are mostly dependent on the programmable access typically realised using APIs and downloadable datasets.

[4]There is not a single authoritative definition of an eResearcher. In this chapter, we consider an eResearcher to be a researcher applying information technology with the goal to analyse or improve the research process. An example might be a researcher applying text-mining to semantically enrich research manuscripts, a person analysing bibliometric publication data or a social scientist looking for patterns in the way researchers collaborate.

Levels of information access	**What does it provide**	**Users group**
Programmable (raw) data access	Access to the raw metadata and content as downloadable files or through an API. The content and metadata might be cleaned, harmonised, preprocessed and enriched.	Developers, digital libraries, eResearchers, companies developing SW, ...
Transaction information access	Access to information primarily with the goal to find and explore content of interest typically realised through the use of a web portal and its search and exploratory tools.	Researchers, students, life-long learners, general public, ...
Analytical information access	Access to statistical information at the collection or sub-collection level often realised through the use of tables or charts.	Funders, government, business intelligence, repository/digital library managers ...

Table 8.2: Types of information communicated to users at the level of granularity they expect - access levels.

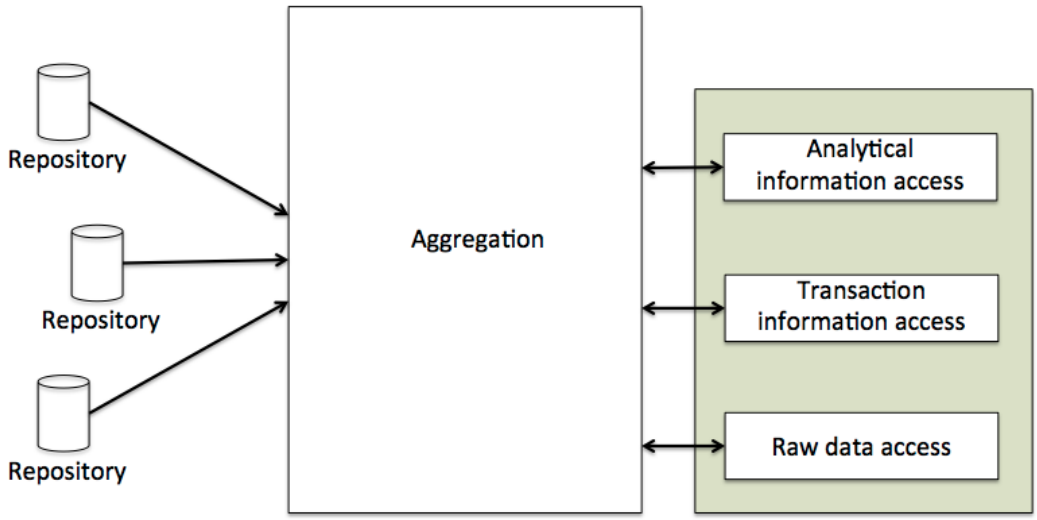


Figure 8.2: The inputs and outputs of an aggregation system.

Figure 8.2 depicts the inputs and outputs of an aggregation system showing the three *access levels*. The internals of the aggregation system are described in the next section. Based on the access level requirements for the individual user groups, we can specify services needed for their support. In Section 8.2, we have discussed that existing OA aggregation systems focus on providing access at one or more of these levels. While together they cover all the three access levels, none of them individually supports all access levels. The central question is whether is it sufficient to build an OA infrastructure as a set of complementary services? Each of these services would support a specific access level and together they would support all of them. An alternative solution would be a single system providing support for all access levels.

One can argue that out of the three access levels, the most essential one is the programmable access level, as all the other levels can be developed on top of this one. This suggests that the overall OA infrastructure can be composed of many systems and services. So, why does the current infrastructure provide insufficient support for these access levels?

While all the needed functionality can be built on top of the first access level, the current support for this level is very limited. In fact, there is currently no aggregation of all OA materials that would provide harmonised, unrestricted and convenient access to OA metadata and content. Instead, we have many aggregations each of which is supporting a specific access level or a user group, but most of which are essentially relying on different data sets. As a result, it is not possible to provide exploratory search on top of the data and make large scale analysis of the OA data. In addition, it is very difficult for developers to improve technology for the upper access levels when their level of access to OA content is limited. From the perspective of link discovery, without the programmable access level, it is not possible to apply link discovery methods, without the transaction level, it is not possible to communicate these links to users (provide exploratory search), and without the analytical level, it is not possible to draw conclusions based on the discovered relationships.

To exploit the opportunities OA content offers, OA technical infrastructure must support all of the listed access levels. This can be realised by many

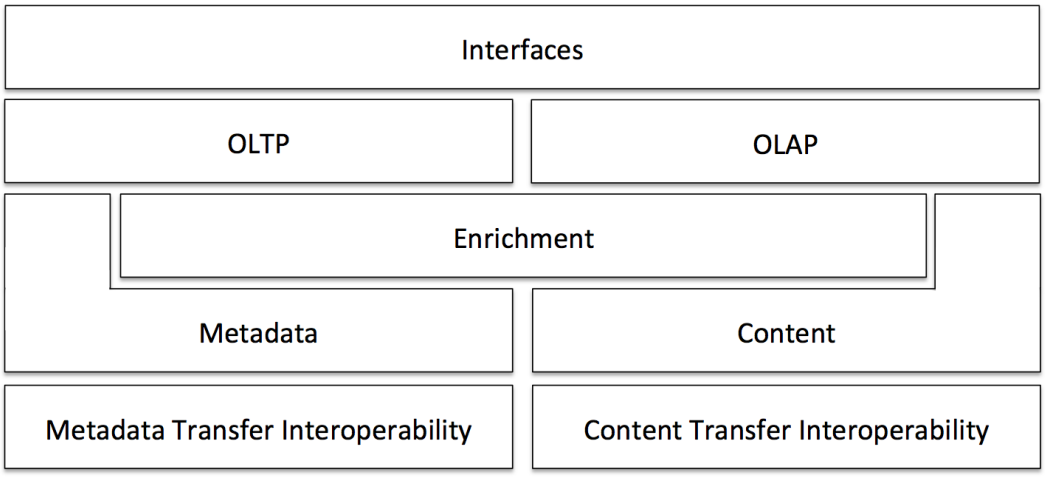


Figure 8.3: Layers of an aggregation system

systems and services, but it is essential that they operate over the same dataset.

8.3.1 The layered model of an aggregation system

We introduce a layered model of the components of an aggregation system that should support the access levels described in the previous section. Figure 8.3 illustrates the hierarchy of the layers in this model. Each layer uses the functionality or resources provided by a lower layer and provides functionality or resources to an upper layer. Each layer works with information at a different level of abstraction. All horizontal layers must be implemented to build an aggregation system and the concrete implementation of these layers in a real system will significantly influence the overall solution. A decision on how to implement a lower layer will affect the functionality of all upper layers.

The Metadata and Content Transfer Interoperability Layer – This layer

implements the necessary processes for metadata harvesting from repositories. Typically, this layer makes use of the OAI-PMH protocol (Initiative, 2002), which is supported by a large proportion of repositories. Content is currently mostly harvested from repositories in non-standard ways, with implementations that are often system specific. This is mainly due to a wide range of practices used in repositories for referencing the content and restrictions on allowing machine access to it, as discussed in (Knoth, 2013). It is expected that, in the future, repositories will support a more sophisticated metadata and content synchronisation mechanism based on ResourceSync (Klein et al., 2013).

The Metadata and Content layer - This layer consists of the metadata and content components. It provides the necessary processes for storing, updating and accessing both content and metadata. The content component is also responsible for various format conversions. The metadata component works with data objects typically represented using technologies, such as XML or RDF, and often conforming to a standard, such as Dublin Core.[5] The content component works with data objects in various open or proprietary formats including the Portable Document Format or the MS Word format.

The Enrichment layer – This layer includes processes for the cleaning and harmonisation of metadata as well as semantic enrichment of the metadata

[5]Although Dublin Core is considered a standard, in practise, Dublin Core metadata provided by different systems differ significantly and consequently there is relatively limited interoperability between systems.

using content-based analysis. Such processes might be fully automated, semi-automated, can make use of crowdsourcing, or can be manual. The approaches can make use of external knowledge sources, such as Linked Data, or various TDM software components, including link discovery. The Enrichment layer can be implemented even if only metadata or only content are available, but this will obviously restrict the functionality that can be provided by the aggregation.

The OLTP and OLAP layer – The layer relies on the input of the Metadata, Content and Enrichment components at least one of which must be implemented by an aggregation system. The layer contains two components, the OLTP and the OLAP component. The OLTP (Online Transaction Processing) component provides the entire functionality necessary for handling transaction-oriented requests. In the context of aggregations of research papers, we can consider transactions as processes providing access and information at the level of a single article or a relatively small set of articles. On the other hand, the OLAP (Online Analytical Processing) component provides the necessary processes for supporting the analysis of the metadata and content held in the aggregation at the level of the whole collection or large sets of items.

The Interfaces layer – The interface layer is responsible for handling and supporting the communication with the users of the aggregation at the desired

level of granularity.

8.3.2 The metadata and content components

Every aggregation system will provide an implementation of the mentioned layers. However the implementation can differ significantly from system to system. A system might choose not to implement a certain component of a layer, but it has to implement all mentioned layers. For example, an aggregation system might choose not to implement the content component, but in that case it has to implement at least the metadata component. Similarly, a system has to implement OLTP or OLAP. Decisions about the extent of the implementation have a high impact on all the upper layers. Therefore, this is particularly important in the case of the *Metadata and Content layer*. If a system does not implement the content component, this will impact all the upper layers. While it will still be possible to perform enrichment or generate statistics, the enrichment will not be able to make any use of text-mining or other content analysis tools and the information provided by the statistics will be limited to that present in the metadata.

Surprisingly, it is possible to see that a large majority of existing OA aggregation systems rely on the use of metadata including BASE and OAISter. Relying purely on metadata has a number of disadvantages among the most important of which are that (1) certain types of metadata can only be created

at the level of aggregation, (2) (pre)-processing can be more effective and efficient at the level of aggregation and finally (3) the availability, validity and quality of aggregated data is typically lower when content cannot be accessed. We will now briefly discuss each of them.

Certain types of metadata can only be created at the level of aggregation. — The content constituting full-texts can be used to extract metadata that cannot be discovered or curated by the data providers themselves. Certain metadata types can hardly be created at the level of the data provider, because they need context or because they are changing in time. For example, metadata describing semantic relatedness of a full-text paper to other papers across repositories cannot be created by the repositories themselves and may change as new content is added to the collection.

(Pre)-processing can be more effective and efficient at the level of aggregation. — Further processing of content supplied by repository providers and its conversion into a more usable format can often be more easily done at the level of aggregation than at the level of the content consumers or providers. Conversions or enrichment processes might require considerable computing resources which might not be available in a small institutional repository, but may be readily available and optimised at the aggregation level. For example, for the purposes of text analysis, access to the plain text is essential. Aggregators can provide the infrastructure for format conversions and provide researchers and

various services with a more convenient way of accessing the data.

Availability, validity and quality — Without access to content, the aggregator has no means for checking content *availability*, *validity* and *quality* as it has to rely on the information provided in the metadata. By availability we understand that the resource described by the metadata record is held in the system and that it is currently available and accessible for a particular audience. Validity means that the metadata used for the description are appropriate. Finally, quality refers to the fact that the resource satisfies the criteria for being offered by the aggregator.

8.4 The CORE system

In this section, we introduce the CORE system and describe the functionality it offers at each of the layers defined in the previous section. The system design is based on the acknowledgement of the issues hindering the impact of OA discussed in the previous sections:

- Support for three groups of users according to the way they mostly communicate with the aggregation (i.e. programmable data access, transaction access, analytical access).

- Open exposure of metadata and content to allow effective reuse and exploitation of information by innovative services.

- Importance of content as a key component of an aggregation system (as opposed to metadata only aggregations).

The main goal of the CORE system is to deliver a platform that can aggregate research papers from multiple sources and provide a wide range of services on top of this aggregation, including those based on link discovery. The vision of CORE is to use this platform to establish the technical infrastructure for Open Access content taking into account the different needs of users participating in the OA ecosystem.

The CORE system offers technical support for three phases: *metadata and full-text content aggregation* (corresponding to the Metadata Transfer Interoperability layer and the Metadata and Content layer), *information processing and enrichment* (corresponding to the Enrichment layer) and, *information exposure* (corresponding to the OLTP and OLAP layer and the Interface layer).

8.4.1 Metadata and full-text content aggregation

In the *metadata and full-text content aggregation* phase, the CORE system harvests metadata records and the associated full-text content from Open Access repositories listed in CORE. The harvesting of the metadata is performed using OAI-PMH requests sent to the repositories.[6] Successful requests return

[6]The CORE system is not inherently limited to any specific harvesting protocol and it enables also other types of data ingestion, such as import from a specific folder in the file system or ingestion of content by crawling content available on the web. However OAI-PMH harvesting dominates over other types of metadata gathering.

an XML document containing information about the papers stored in a repository. A good practice in repositories, unfortunately not consistently applied, is to provide as part of the metadata the links to the full-text documents.[7] The CORE system extracts these links and uses them to download full-texts from repositories. This process is, in reality, more complicated as some repository systems do not expose the link to the full-text, but only a link to a page close to the full-text. This issue is discussed in detail in (Knoth, 2013). The CORE system therefore uses a notion of *harvesting levels* to crawl from a given URL into certain depth to discover the full-text to be downloaded. The system then carries out format conversions, such as the extraction of plain text.

The CORE system supports the harvesting and downloading of content from multiple repositories at the same time and has been optimised to utilise architectures with multiple processors. The harvesting component in CORE can be controlled using a web interface accessible to the system administrator. The system supports adding, removing and editing repositories, importing and synchronising repository updates with repository registries, such as the OpenDOAR system[8], and the scheduling and monitoring of the harvesting process.

[7]The OAI-PMH protocol itself is not directly concerned with the downloading of full-text content, as it focuses only on the description and the transfer of metadata.

[8]http://www.opendoar.org/

8.4.2 Information processing and semantic enrichment

The goal of the *information processing and semantic enrichment* is to harmonise and enrich the metadata using the harvested metadata itself, the full-text content and information from external systems. Given the fact that metadata harmonisation and cleaning in aggregation systems is the de facto standard, we will focus on how the CORE system utilises the full-text. First, the system runs a standard text preprocessing pipeline including tokenisation, filtering, stemming and indexing of the metadata and text. A number of text mining tasks are then performed. They include primarily the *discovery of links to semantically related content*, but also other tasks, such as the *categorisation of content* and the *extraction of citations and citation resolution*, which we mention here, but are currently not the focus of our attention. Naturally, new text and data mining algorithms can be added to CORE (or plugged into it using the CORE API) and we plan to do so as we progress.

Discovery of links to semantically related content — semantically related content is identified and the information about the measured semantic similarity is then used for a number of purposes, such as content recommendation/navigation and duplicates detection. The system supports the recommendation of full-text documents related to a metadata record and the recommendation of a semantically related item held in the aggregator for an arbitrary resource on the Web. We describe in detail the way semantic similarity is

calculated in CORE in Section 8.4.3 and discuss the use of this information in Section 8.4.6.

Categorisation of content — content stored in OA repositories and journals reflects the diversity of research disciplines. Information about the specific subject of a paper (e.g. computer science, humanities) can be, for example, used to narrow down search, to monitor trends and to estimate content growth in specific disciplines. Only about 1.4% (Pieper and Summann, 2006) of items in OA repositories have been classified according to some taxonomy and manual classification is costly.

Extraction of citations and citation resolution — CORE uses the ParsCit system[9] to extract citation information from the publications full-text and the resource metadata and CrossRef API to acquire DOI identifiers for the cited references. This information is used in turn to check if the (cited) target documents are also present in the CORE aggregation to establish a link between the cited publications.

8.4.3 Link discovery in CORE

The CORE system contains a monolingual document-to-document link discovery engine. CORE estimates semantic relatedness between two texts using the cosine similarity measure calculated on term-document vectors where each

[9] http://aye.comp.nus.edu.sg/parsCit/

dimension of the vector (term) is weighted using the $tfidf$ schema, i.e. this is the same approach as the one used in Chapter 4. Due to the size of the CORE dataset[10] and consequently the high number of combinations, semantic similarity cannot be calculated for all document pairs in a reasonable time. To make the solution scalable, CORE uses a heuristic, to decide which document pairs are unlikely to be similar and can therefore be discarded. This allows CORE to cut down the amount of combinations and to scale up the calculation to millions of documents.

We will now have a look on how this technique works. The idea is similar to the *df*-cut approach tested by Elsayed et al. (2008) and previously introduced in Section 3.2.1.1. In order to find the most similar document to another document in a document collection, one needs to calculate the similarity of this document to all other documents. This leads to a high number of combinations. However, in practice, we only need to consider all documents that share at least one word with the document we want to evaluate. Finding these documents is simple when the inverted index structure is available. It is the union of all documents that appear in the posting lists for all terms in the document we are evaluating. The problem is that a few common words will be typically responsible for a situation in which every document shares a few words with all other documents in the document collection. Consequently, the *df*-cut approach is motivated by the fact that removing a small set of the most common

[10]Currently over 20 million records from over 600 repositories

terms (for example 1%) from the inverted index greatly reduces the number of combinations that need to be taken into account. The *df*-cut approach (see Figure 8.4) can be applied directly on the inverted index structure, prior to considering any particular document, by removing a set of rows representing the most common terms in the whole collection.

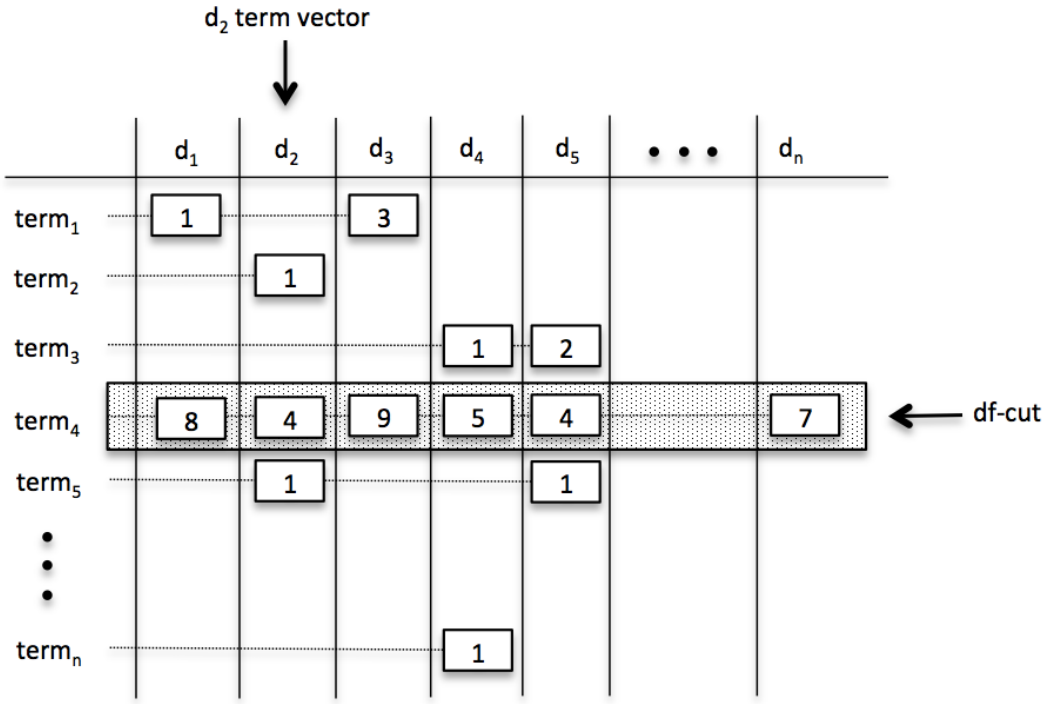


Figure 8.4: The *df*- and *tfidf-cut* approach. Terms appearing in many documents have a high document frequency (low *idf*). By not considering $term_4$ in finding documents similar to d_2, we only need to consider d_5 instead of all documents in the collection, saving a significant amount of computing resources.

The approach implemented in CORE is based on the idea of a $tfidf$-cut, which is applied for each document we want to calculate similarity for separately. The rational is that words with low $tfidf$ are unlikely to sufficiently

contribute to the final value of the similarity with other documents, while they might be causing the explosion in the number of combinations. In practical terms, the $tfidf$-cut is achieved by retrieving the term vector for a specific document and then by removing all terms in it with $tfidf$ lower than a certain threshold. Providing that this threshold is not very high, it is unlikely this procedure would cause missing the identification of highly similar documents. In terms of time complexity, the only difference between the df-cut and $tfidf$-cut approaches is the need to filter words according to a specific $tfidf$ threshold at the time a term vector is retrieved from the inverted index. For a large document collection, this will only cost constant time per each document the similarity of which we are calculating. Therefore, calculating pair-wise similarity should, based on the results of Elsayed et al. (2008), empirically scale approximately linearly with both the application of the df-cut and $tfidf$-cut. This also means the time needed for the calculation of a set of n most similar documents to a given a document d can be achieved in a constant time with respect to the number of documents in the collection.

In addition to the $tfidf$-cut and motivated by the results presented in Chapter 5, the CORE systems uses an empirically determined threshold of 0.95 to identify duplicate content. Even though the detection of related content is fairly quick taking on average 1-3 seconds for a given document in the current infrastructure, it would not be practical to recalculate this every time

this information is needed. Up to 10 generated links with similarity higher than 0.3 are therefore cached in the database and made readily available for quick retrieval. We make use of the fact that the similarity relationship is symmetric in order to modify cached information if a link from a newly added document points to another document for which we already have a cached version. Since the CORE corpus changes as it grows, this can also slightly influence the word frequency statistics (in particular the $tfidf$ scoring), which are used to determine the similarity values. To face this issue, CORE forces the recalculation of cached links after a set time period elapses.

8.4.4 Validating link discovery in CORE using citation information

One of the challenges we faced with the CORE link discovery system was to validate the assumption that the value of semantic similarity can be used to identify links in the same way as described in Chapter 4. More specifically, we were interested whether the observed relationship between semantic similarity and linked-pair likelihood measured on the Wikipedia collection and shown in Figure 4.2 can be generalised to the collection of research papers.

As research papers are traditionally created as PDF files (not hypertext), they typically do not contain the same types of links as those found in Wikipedia. While Wikipedia links usually connect an anchor with a page describing a con-

cept referenced by the anchor, research papers traditionally use citations to point to relevant related work. As described by Garfield et al. (1965), citations are created for a wide range of reasons, including giving credit for related work, identifying methodology, equipment, and the like, providing background reading, correcting existing work, criticising previous work, substantiating claims, alerting researchers to forthcoming work and arguing against the work or ideas of others. Thus, it is a question whether the value of semantic similarity is correlated with the appearance of a citation link connecting two research papers in the same way as it is the case in Wikipedia.

As the CORE system has access to fulltexts of research papers as well as link discovery and citation extraction modules, this makes it possible to investigate the above mentioned relationship. We have extracted from CORE a sample of citations for which there is a fulltext in CORE for both the cited as well as the citing document. This resulted in a dataset of 18,460 citations with full-texts from a wide range of disciplines. We then measured the semantic similarity of the cited and citing fulltexts and created a histogram showing the similarity of these citation pairs (Figure 8.5). It should be noted that the figure does not show linked-pair likelihood, which would be difficult to estimate due to the sparsity of citation information. More specifically, as citations can point to any other (even non-OA) document, only a small fraction of citations reference documents inside the CORE collection. Calculating linked-

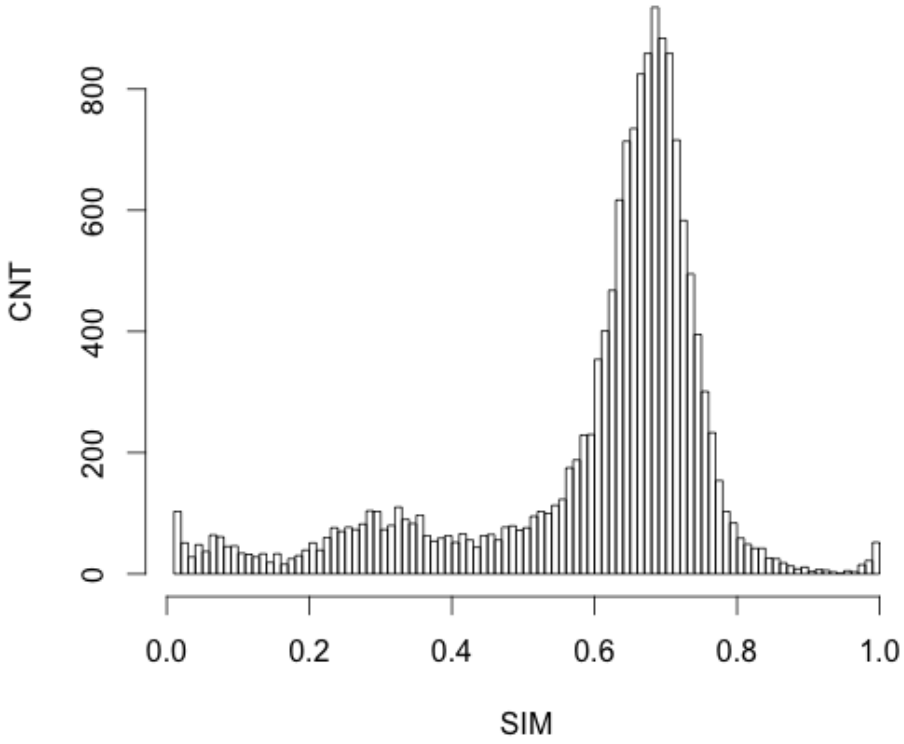


Figure 8.5: The number of citation pairs (CNT) and their similarity (SIM) calculated by the CORE link discovery system. The histogram has been produced on a random sample of 18,460 citation pairs from CORE with fulltext.

pair likelihood would require carrying out the experiment with unlinked pairs as well. This would be difficult to do with the existing dataset.

The histogram in Figure 8.5 is interesting for two reasons. Firstly, it shows that the distribution of the number of citations with respect to semantic similarity resembles the normal distribution. Assuming that the similarity distribution of all document pairs (including those that are not linked using a

citation) follows a power law distribution, as in Figure 4.1, we can imply that it is reasonable to use the value of semantic similarity as an indicator for linking research papers. Secondly, the peak of the histogram in Figure 8.5 is at similarity of about 0.68. Comparing this with Figure 4.2 indicates that the peak appears to be fairly close, though not identical, to the one measured on the Wikipedia collection. While it is not possible to draw conclusions from the direct comparison of these two figures (as Figure 8.5 does not take into account unlinked document pairs), the findings are consistent with the results reported in Chapter 4.

8.4.5 Information exposure

In the *information exposure* phase, the CORE system provides a range of services for accessing and exposing the aggregated and enriched data. At the moment, the services are delivered through the following applications: *CORE Portal*, *CORE Mobile*, *CORE Plugin*, *CORE API*, *CORE Data Dumps* and *Repository Analytics*. All of these tools utilise the information provided by the lower layers, in particular the Semantic Enrichment layer and the Metadata and Content layer.

CORE Portal[11] – is a web-based portal for searching, exploring and accessing the aggregated content. CORE Portal is not just another search engine for scholarly materials. CORE follows the idea that a resource can only be

[11] http://core.kmi.open.ac.uk

regarded as Open Access if its full-text is openly accessible. While this might sound trivial, the usual practice of Open Access aggregators is to aggregate metadata of resources and not to check the availability of the full-text. CORE takes a different approach ensuring the availability of information specified in the metadata. Consequently, all search results produced by the system as a response to a user's query will contain links to openly accessible full-texts (unless a metadata search is explicitly requested by the user), for the purposes of availability and reliability cached on the CORE server. In addition to search, the CORE Portal offers other services on top of the aggregated Open Access collection utilising the information provided by the lower layers, including the use of discovered links for duplicates detection and navigation (further discussed in Section 8.4.6) and citation extraction.

CORE Mobile[12] [13] – offers pretty much the same functionality as the CORE Portal, but has been developed for a mobile application. It is an application for iOS (iPhone, iPad, iPod Touch) and Android devices, which can be used on both smartphones and tablet devices. The application provides search and navigation capabilities across related papers stored in OA repositories. It also supports downloading of full-text articles to the mobile devices.

CORE Plugin[14] – A platform- and browser-independent plugin for digital libraries and institutional repositories that exposes discovered links to related

[12]https://play.google.com/store/apps/details?id=uk.ac.open.core.mobile
[13]http://itunes.apple.com/lk/app/core-research-mobile/id523562663
[14]http://core.kmi.open.ac.uk/intro/plugin

documents. The plugin recommends semantically related papers to the document currently being visited and the recommendations are based on either full-text or metadata.

CORE API[15] – In fact, CORE offers two APIs enabling external systems and services to interact with the CORE system. The first is a fairly popular RESTful API which supports tasks, such as searching for content using various criteria, downloading PDF or plain text documents, getting information about related documents and detecting the subject of a research text. The API communicates using RDF or JSON. The second API is a SPARQL endpoint to the CORE repository[16] registered in the Linked Open Data cloud. It provides information about the harvested papers and their similarities encoded in the RDF format. An example using the dataset schema to describe an article of interest is depicted in Figure 8.6.

CORE Data Dumps[17] – Provide a downloadable package of all the aggregated data and content in a JSON formatted file and with the semantic enrichments, for example, in the form of the discovered links. The full-text is available as plain text. This dataset is particularly useful for researchers and can be used as a input data for experiments.

Repository Analytics[18] – A tool that enables to monitor the ingestion of

[15]http://core.kmi.open.ac.uk/intro/api
[16]http://thedatahub.org/dataset/core
[17]http://core.kmi.open.ac.uk/intro/data_dumps
[18]http://core.kmi.open.ac.uk/repository_analytics

metadata and content from repositories and provides a range of data and usage statistics including the amount of content, accesses to the content, availability and validity of the metadata and content. The aim is to facilitate the process of increasing the interoperability of repositories.

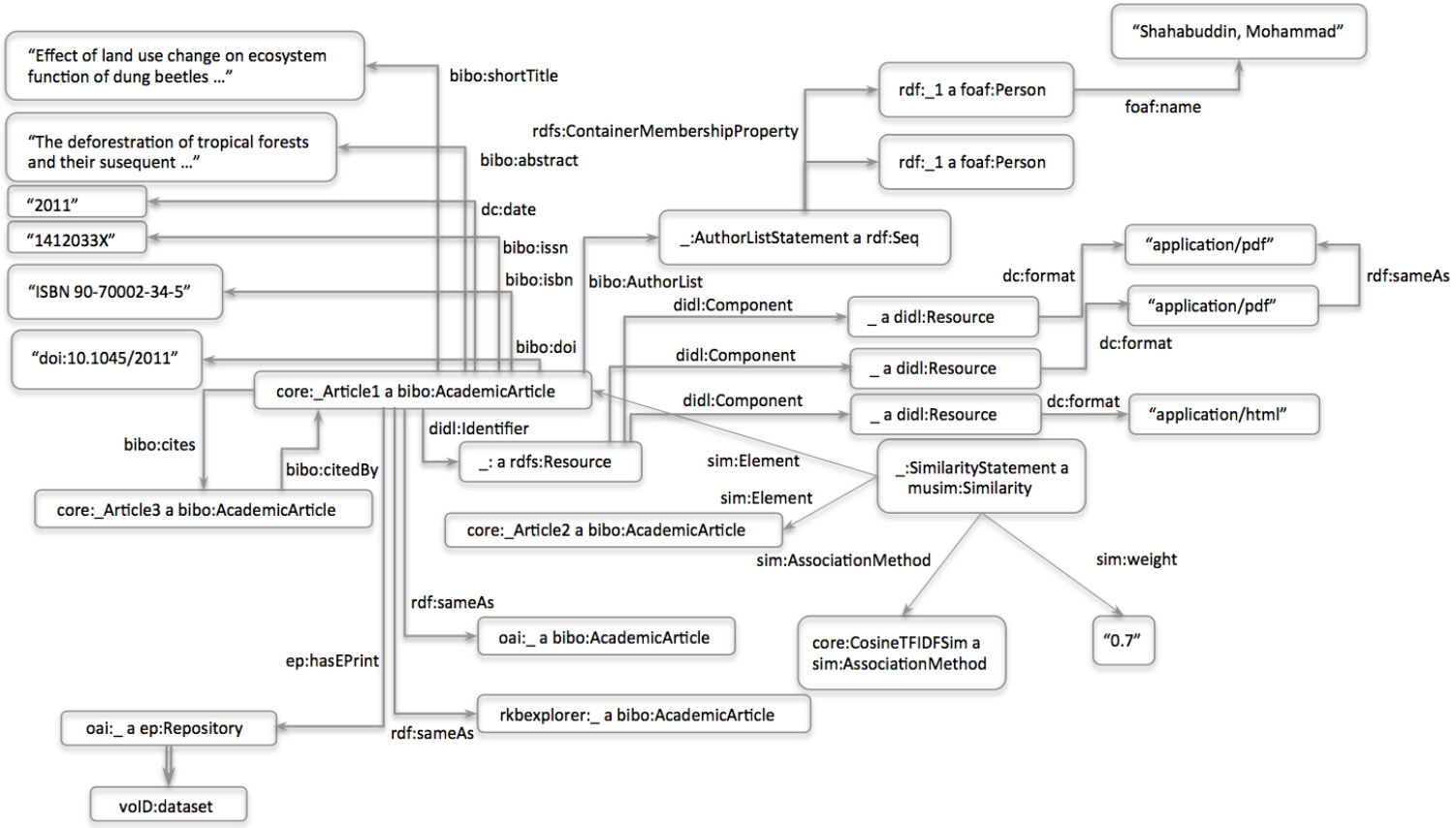


Figure 8.6: An example schema demonstrating how data are represented in the CORE data dumps. The representation uses vocabulary from a number of ontologies. The information about the discovered links is encoded using the vocabulary from the Music Similarity Ontology (MuSim) (Jacobson et al., 2010).

8.4.6 Using discovered links to support exploratory search

While we have already described in detail how links are generated by CORE, the eventual improvement to the accessibility of content in the document collection happens typically at the level of the user interface[19]. The CORE system exposes information about the generated links on the CORE Portal and using the CORE Plugin. Additionally, we have designed a prototype of a visual exploratory search interface, which makes use of the generated links.

8.4.6.1 Presenting recommendation links on the CORE Portal

The integration of the information about the discovered links is presented on the CORE Portal either as a list (Figure 8.7), which is the default setting, or using a graph view (Figure 8.8). The information about related resources to a reference resource is available in the resource's details page. We have deliberately used, as an example, one of the resources that resulted from the research presented in this thesis to demonstrate the qualitative nature of the recommendations. The paper to which the discovered links are displayed in Figure 8.7 presents the results of the methods developed for NTCIR-9 discussed in Chapter 7 of this thesis. We can see that the first recommendation displayed by CORE is a paper which has resulted from the research presented

[19]Unless machine to machine interaction through an API/data dump is desirable.

in Chapter 6.[20] Since it makes sense to read Chapter 7 immediately after Chapter 6, this is a sound recommendation. The second best recommendation is the NTCIR-10 work presented also in Chapter 7. Again a very good recommendation. The third most recommended paper identified by CORE is actually also the result of research described in this thesis and refers to the topic presented in Chapter 4. By manually visiting all the recommendations, we can see that they are all relevant to the topic.

The CORE Portal also informs about identified duplicate items by creating a shared page for these resources. For example, it has been identified that an article shown in Figure 8.9 is available from four repositories. It should be noted that the identification was successful although the PDF documents were not exactly the same. For example, the document in the Open Research Online repository has a cover page, while in the Southampton repository it does not. The settings for the threshold to decide whether an item is related or duplicate followed our research described in Chapters 4 and 5.

8.4.6.2 Presenting recommendation links using the CORE Plugin

The CORE Plugin enables the embedding of the CORE link discovery system in third party systems. In this case, the reference resource does not have to be stored in the CORE system for CORE to be able to supply the recom-

[20]The algorithm used in CORE for discovering links does not boost the recommendations of articles that share an author, thus the articles are recommended due to the semantic similarity of their content.

mendations. The system on which the plugin is deployed only sends a request against the CORE API with all necessary information. The CORE Plugin then displays the recommendations. From a user interface perspective, the plugin can be integrated to a third party system in a way that the user is unlikely to realise the links are coming from a third party system.

The CORE Plugin has been integrated into the systems of various repositories, such as Open Research Online (Figure 8.10), and library systems, such as the portal the European Library (Figure 8.11). The plugin is also available from the EPrints Bazaar, which makes it very easy to integrate CORE into any EPrints Repository. There is also a version of the plugin for Open Journal Systems (OJS), which is a leading platform used by OA journal publishers developed by the Public Knowledge Project[21].

8.4.6.3 Using visualisations as a means for exploring document collections

[21] `https://pkp.sfu.ca/`

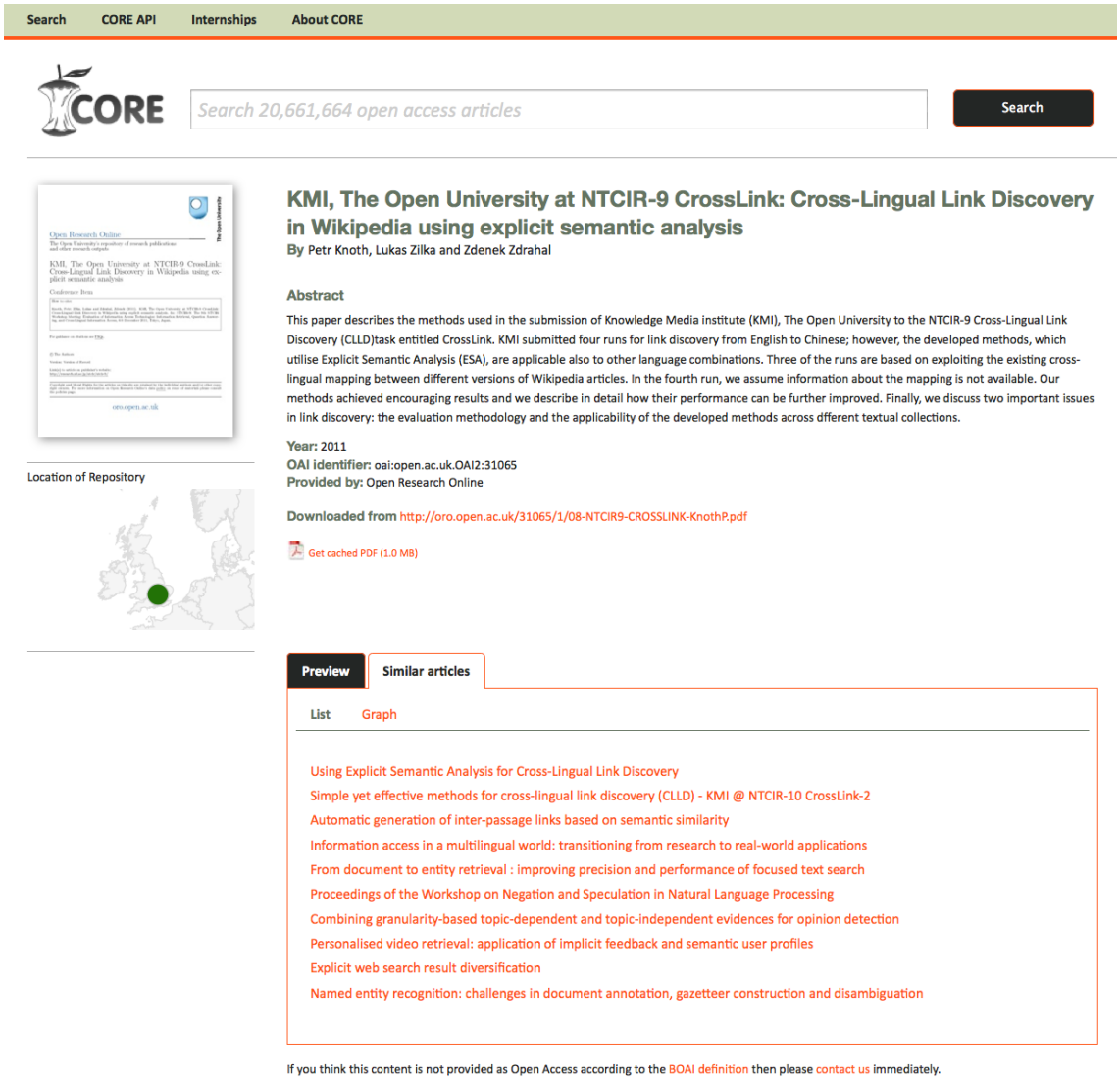


Figure 8.7: Presenting the document-to-document generated recommendation links on the CORE Portal as an ordered list.

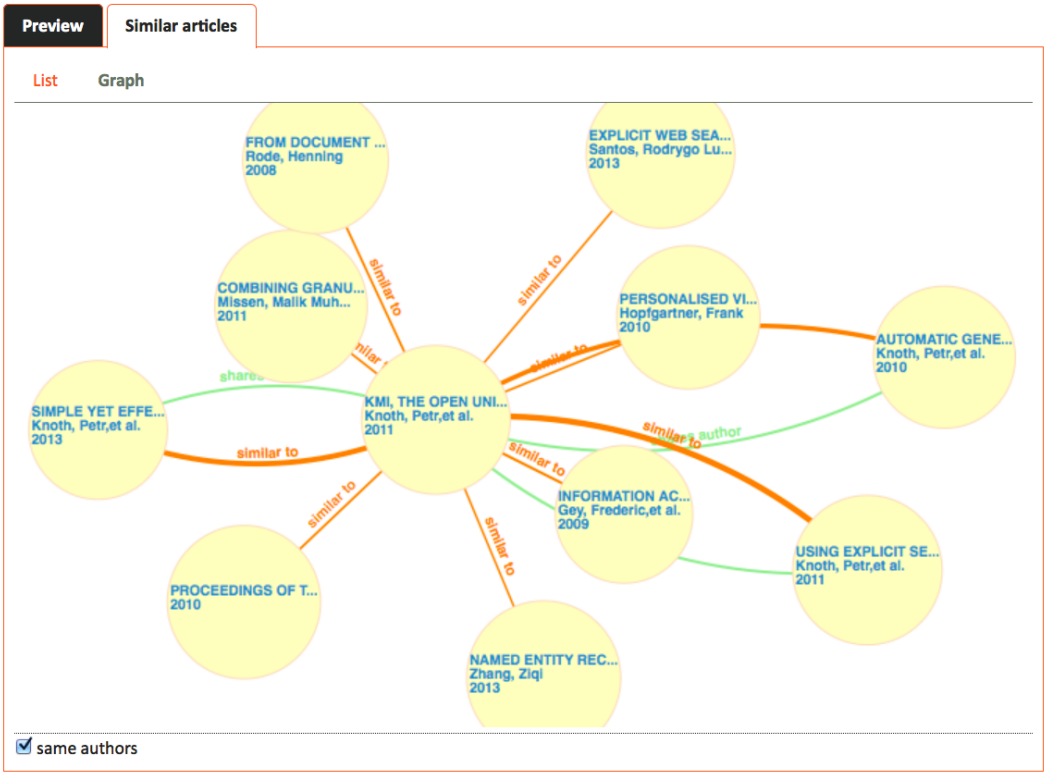


Figure 8.8: Presenting the document-to-document generated links on the CORE Portal using a graph view. Articles that share an author are highlighted.

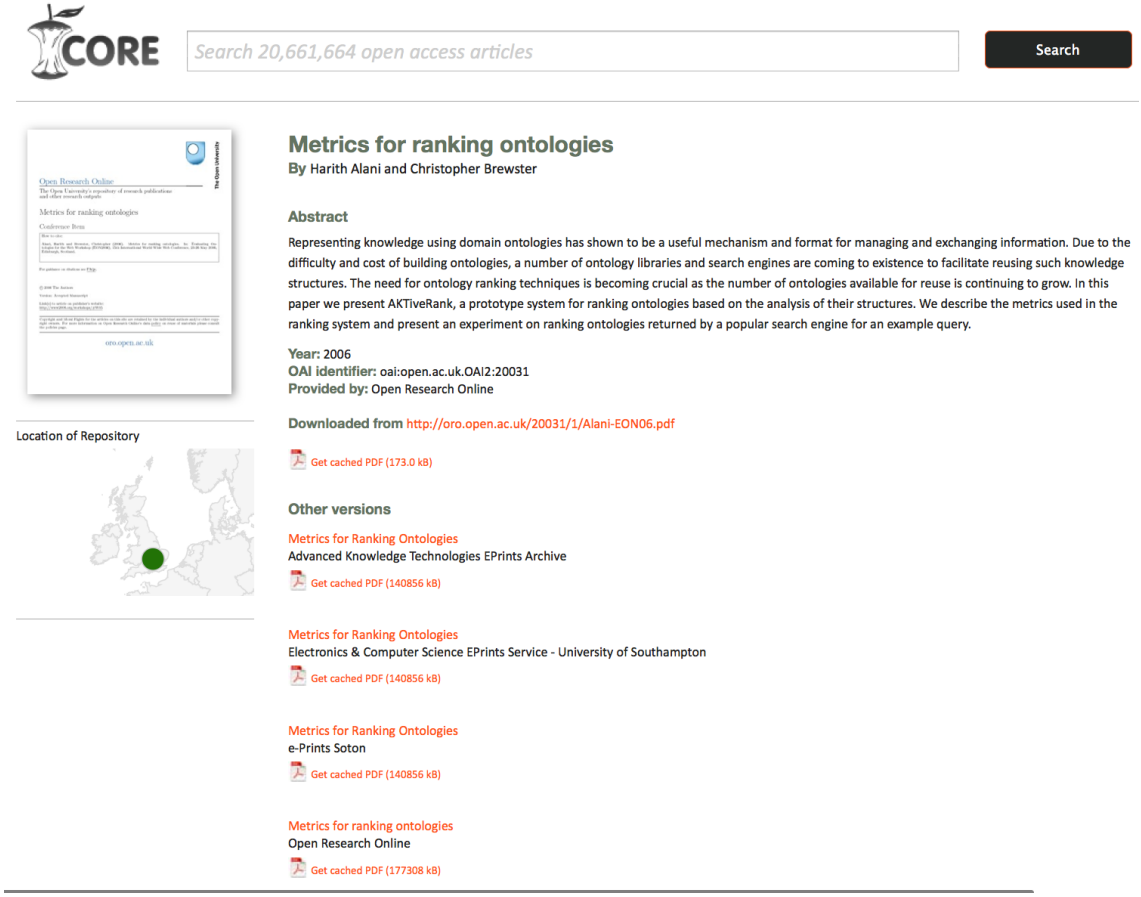
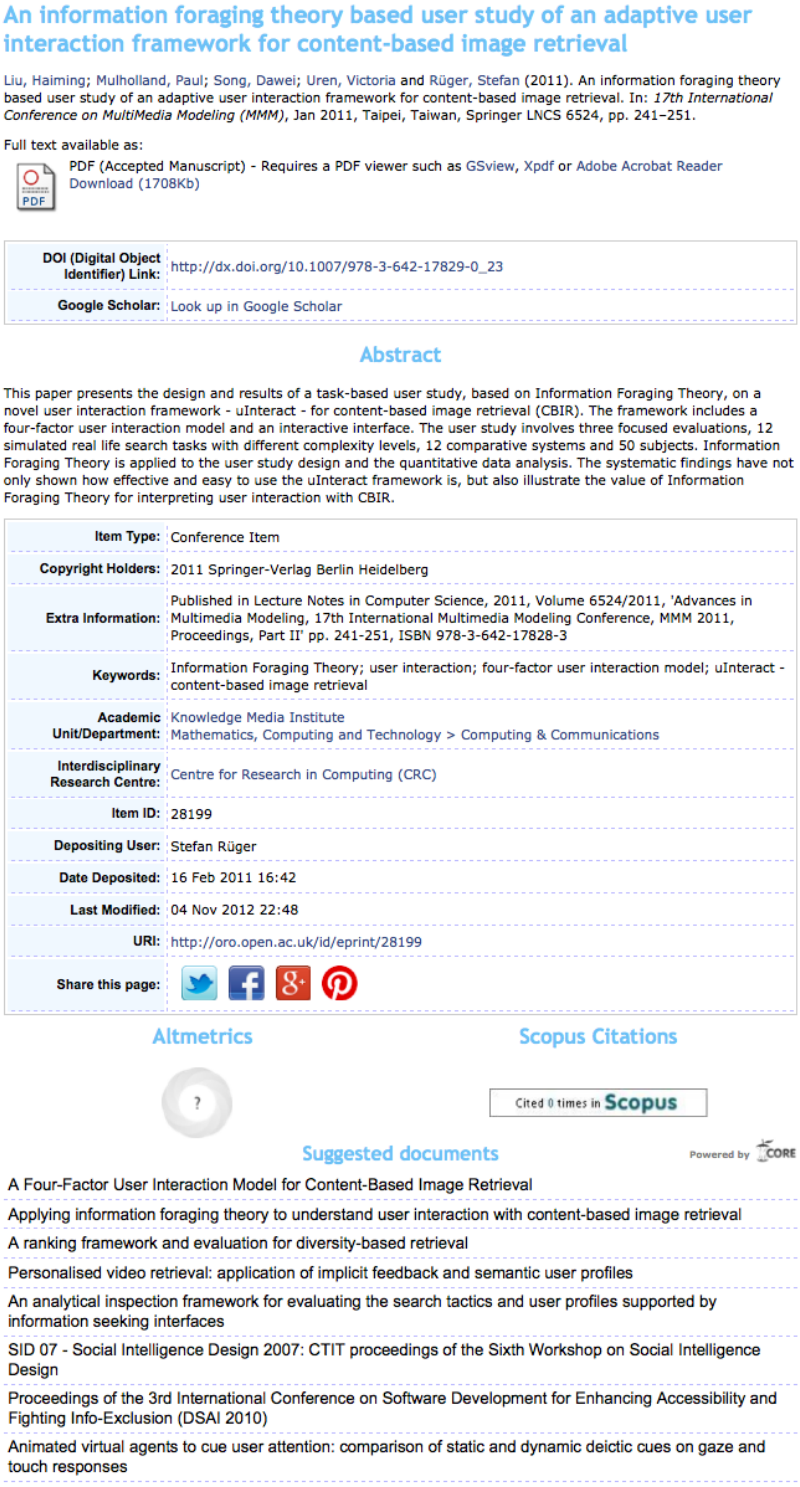


Figure 8.9: Presenting duplicate items on the CORE Portal.



An information foraging theory based user study of an adaptive user interaction framework for content-based image retrieval

Liu, Haiming; Mulholland, Paul; Song, Dawei; Uren, Victoria and Rüger, Stefan (2011). An information foraging theory based user study of an adaptive user interaction framework for content-based image retrieval. In: *17th International Conference on MultiMedia Modeling (MMM)*, Jan 2011, Taipei, Taiwan, Springer LNCS 6524, pp. 241–251.

Full text available as:

PDF (Accepted Manuscript) - Requires a PDF viewer such as GSview, Xpdf or Adobe Acrobat Reader
Download (1708Kb)

DOI (Digital Object Identifier) Link:	http://dx.doi.org/10.1007/978-3-642-17829-0_23
Google Scholar:	Look up in Google Scholar

Abstract

This paper presents the design and results of a task-based user study, based on Information Foraging Theory, on a novel user interaction framework - uInteract - for content-based image retrieval (CBIR). The framework includes a four-factor user interaction model and an interactive interface. The user study involves three focused evaluations, 12 simulated real life search tasks with different complexity levels, 12 comparative systems and 50 subjects. Information Foraging Theory is applied to the user study design and the quantitative data analysis. The systematic findings have not only shown how effective and easy to use the uInteract framework is, but also illustrate the value of Information Foraging Theory for interpreting user interaction with CBIR.

Item Type:	Conference Item
Copyright Holders:	2011 Springer-Verlag Berlin Heidelberg
Extra Information:	Published in Lecture Notes in Computer Science, 2011, Volume 6524/2011, 'Advances in Multimedia Modeling, 17th International Multimedia Modeling Conference, MMM 2011, Proceedings, Part II' pp. 241-251, ISBN 978-3-642-17828-3
Keywords:	Information Foraging Theory; user interaction; four-factor user interaction model; uInteract - content-based image retrieval
Academic Unit/Department:	Knowledge Media Institute Mathematics, Computing and Technology > Computing & Communications
Interdisciplinary Research Centre:	Centre for Research in Computing (CRC)
Item ID:	28199
Depositing User:	Stefan Rüger
Date Deposited:	16 Feb 2011 16:42
Last Modified:	04 Nov 2012 22:48
URI:	http://oro.open.ac.uk/id/eprint/28199
Share this page:	

Altmetrics

Scopus Citations

Cited 0 times in Scopus

Suggested documents

Powered by CORE

- A Four-Factor User Interaction Model for Content-Based Image Retrieval
- Applying information foraging theory to understand user interaction with content-based image retrieval
- A ranking framework and evaluation for diversity-based retrieval
- Personalised video retrieval: application of implicit feedback and semantic user profiles
- An analytical inspection framework for evaluating the search tactics and user profiles supported by information seeking interfaces
- SID 07 - Social Intelligence Design 2007: CTIT proceedings of the Sixth Workshop on Social Intelligence Design
- Proceedings of the 3rd International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion (DSAI 2010)
- Animated virtual agents to cue user attention: comparison of static and dynamic deictic cues on gaze and touch responses

Figure 8.10: Integration of the CORE Plugin into Open Research Online.

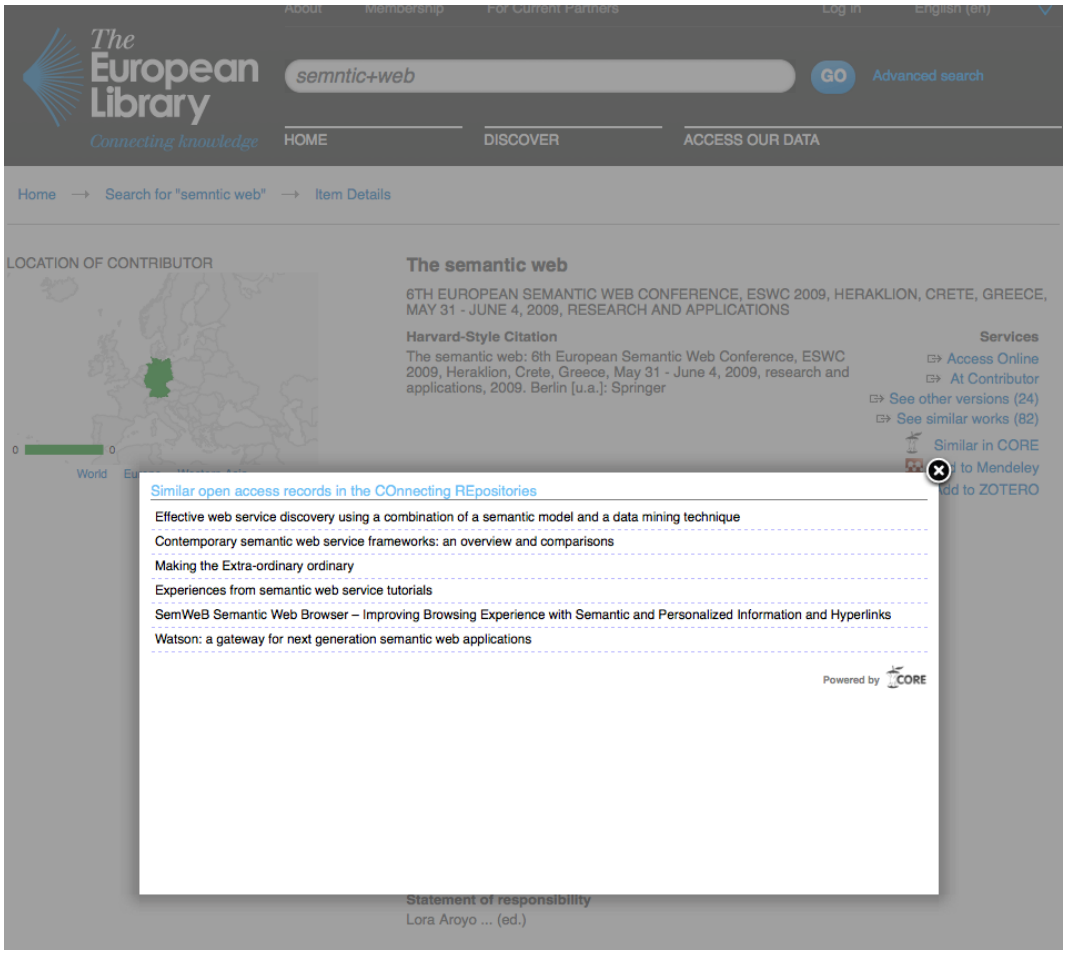


Figure 8.11: Integration of the CORE Plugin into the European Library portal.

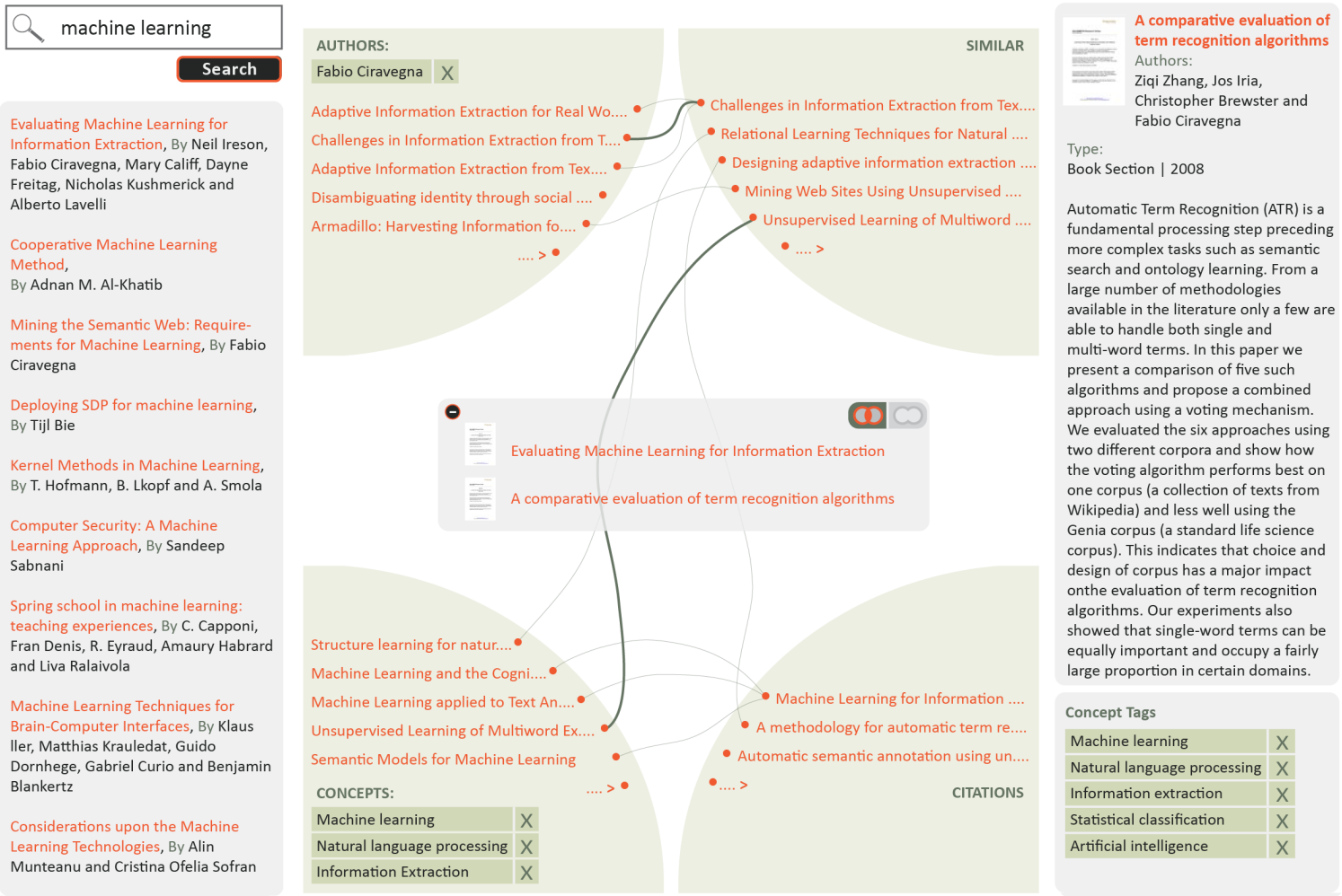


Figure 8.12: The exploratory visual search interface for CORE. The image is taken from (Herrmannova and Knoth, 2012)

As both the link discovery presentation approach on the CORE Portal and in the CORE Plugin can be seen as a fairly standard approach, Herrmannova and Knoth (2012) have also studied how more innovative exploratory search interfaces could be developed with the help of visualisations. While exploratory searches constitute a significant proportion of all searches (Rose and Levinson, 2004), it is interesting to see that current search interfaces typically do not sufficiently support them. Visual search interfaces make use of our spatial skills in order to help us to navigate through content. An important aspect of visualisations is that they make it easier to communicate structure, organisation and relations in content. They can also be well utilised to improve the search experience, by depicting more information than a typical text search interface using the same space, and they can simplify the process of finding relevant information. The visualisation, depicted in Figure 8.12, can be classified as a document level query-focused visualisation. It tries to visualise attributes of the collection items, their mutual links and relations in response to a user supplied query. The visualisation provides support for exploring document relations, discovering interesting connections across dimensions/facets and comparing and contrasting documents. Although the visualisation has been designed to support exploratory search on CORE, it is applicable also to other domains. More details on this visualisation can be found in (Herrmannova and Knoth, 2012).

8.5 Serving the needs of multiple user groups

Since CORE supports all three types of access specified in Table 8.2, and primarily the programmable access layer, it potentially provides functionality for all the user groups identified using a single dataset and at the level of the content (not just metadata). While we do not claim that CORE provides all of the functionality that these user groups need or desire, we claim that this combination provides a healthy environment on top of which the overall OA technical infrastructure can be built. To give an example, it allows researchers accessing the dataset and experimenting with it, for example, to develop new methods for discovering hidden relationships or new trend visualisations. The crucial aspect is that the method can be evaluated with respect to existing services already offered by CORE (or anybody else) built on top of the CORE aggregated dataset, i.e. the researcher has the same level of access to the data as the CORE services. The method can now also be implemented and provided as a service on top of this dataset. The value of such an infrastructure is in the ability to interact with the same data collection at any point in time at the three access levels. This allows the propagation of findings and ideas in a top down or bottom up fashion between these levels, and thereby, between different user groups. This creates an environment in which technologies can be applied and tested soon after they are developed and using a representative

data sample the results can be analysed and fed back.

A question one might ask is why should an aggregation system like CORE provide support for all three access levels when many might see the job of an aggregator as just aggregating and providing access to the content? As we already explained in Section 8.3, the whole OA technical infrastructure can consist of many different services providing that they are built on the same dataset. While CORE aims to support others in building their own applications, we also recognise the needs of different user groups (apart from researchers and developers) and want to support them. While this might seem a dilution of effort, our experience indicates that about 90% of time is spent in aggregating, cleaning and processing data, and only the remaining 10% in providing services on top of this data. This is consistent with the findings of Manghi, Mikulicic, Candela, Castelli and Pagano (2010) who notes that building aggregations is an expensive task. It is therefore not only needed that research papers are Open Access, the OA technical infrastructures and services should also be metaphorically "Open Access." This will bring new opportunities for the development of innovative applications, allowing exploratory and analytical access to the content while at the same time providing all basic functionalities users need, including look-up searching and access to research papers.

8.6 Future work on CORE

At the time of writing, CORE has already aggregated millions of articles from hundreds of OA repositories.[22] In the future, we aim to work towards improving the freshness of information in CORE as well as adding more content and repositories as they appear. CORE uses the APIs of repository registries, such as OpenDOAR[23] to discover new repositories and update information about the existing ones. As OA is quickly becoming the default policy in some countries, we can expect that the majority of research papers will soon become freely available on the Web, further increasing the CORE benefits.

In terms of services, we aim at adding more semantic enrichment processes and making use of their results at all three access levels with a particular focus on the programmable data access level currently realised through the CORE API and CORE Data Dumps. For example, we aim at adding and releasing the citation graph together with the graph of semantically related content, created by link discovery methods, to facilitate the research analysing how research communities co-operate. Such work enables, for instance, experimenting with new evaluation metrics approaches, such as Semantometrics (Knoth and Herrmannova, 2014), and their relationship to traditional Bibliometric measures based on citations. In terms of link discovery, we believe there is a potential

[22] At the time of writing, this is about 20 million metadata records and 2 million fulltexts from about 600 repositories.

[23] http://www.opendoar.org/

to further improve exploratory experiences by developing lower granularity link discovery systems for research papers, by deploying new methods for link typing optimised for research papers and by interlinking the collection of research papers with other supporting databases, such as patents, newspapers and books.

8.7 Discussion

As shown in Figure 8.13, the development of CORE has been motivated by the opportunity to apply link discovery in a large collection of research papers, where exploratory experiences, including discovery, are needed. The resulting system and aggregated data enable further research in link discovery. Moreover, CORE actually further motivates research of a wide range of TDM use cases. This section discusses these wider benefits.

A study into the Value and Benefits of Text Mining commissioned by Jisc in 2012 concluded that text-mining of research outputs offers the potential to provide significant benefits to the economy and the society in the form of increased researcher efficiency; unlocking hidden and developing new knowledge and improving the research process and its evidence base. These benefits will result in significant cost savings and productivity gains, innovative new service development, new business models and new medical treatments (McDonald and Kelly, 2012). Unfortunately, exploiting the potential of text mining has

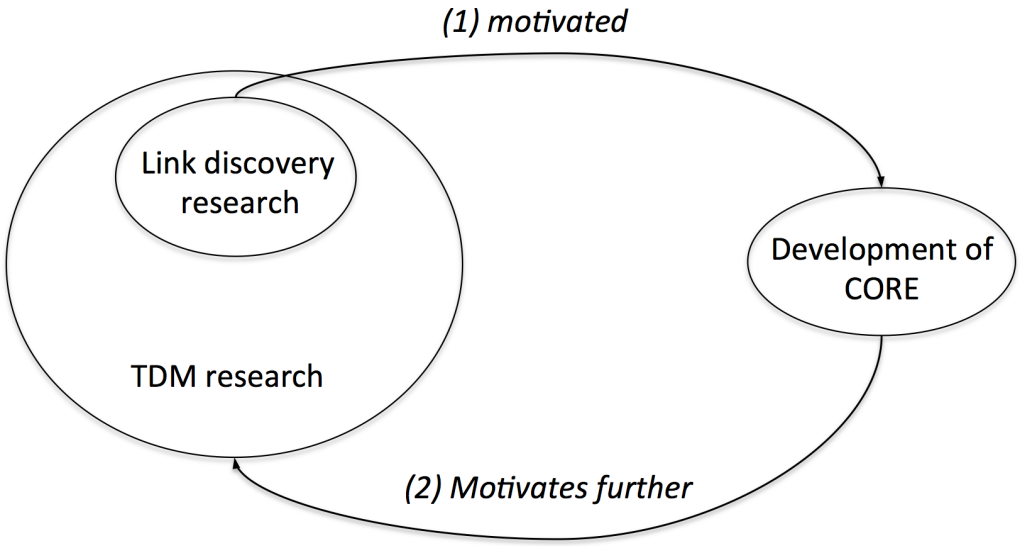


Figure 8.13: Link discovery research presented in this thesis motivated the development of CORE. However, the implications of CORE development motivate further research in TDM, providing benefits beyond link discovery.

for a long time been prevented by a range of legal and technical restrictions. On the legal front, a significant progress has been made recently. A copyright exception for text and data mining for non-commercial and research purposes has been passed by the UK Parliament and came into effect on 1st June, 2014 (*Implementing the Hargreaves review*, 2014). It is now important that similar legislation will follow in other countries.

Many of the technical restrictions on text mining are equally important as the legislation issues and deserve the same degree of attention. These restrictions refer primarily to problems in accessing data from hundreds or even thousands of systems many of which:

- Do not offer machine accessible APIs for full-texts at all or use propri-

etary non-standard solutions often limiting the degree to which text-mining, such as link discovery, can be applied on the dataset.

- Do not technically allow the full transition of the dataset outside of their infrastructure (although legally this is possible), for example due to hardware limitations of the content provider, preventing the ability to process data close to where they are stored, which is a performance-related requirement for many text-mining algorithms.

- Describe metadata and content in different formats or interpretations of these formats causing a lack of metadata interoperability.

- Do not subscribe to a single authentication mechanism that would work across providers, making it very hard for text-mining applications to process data at scale.

The significance of these issues is especially high due to the fact that some data providers (typically commercial publishers) are worried that the increase of interoperability would undermine their business models and are consequently not motivated in lowering these technical barriers. In this sense, where legal restrictions have been lifted, technical barriers to actually gaining access have now become the new battlefield. The position of these organisations is well illustrated on the statement of Richard Mollet, the head of the UK Publishers Association: "I cannot say strongly enough: we support

content mining. It can only work well if we are involved in the process and managing the access" (White et al., 2011). The recent Elsevier text-mining policy demonstrates how technical restrictions widely limiting the potential impact of text-mining (Reilly, 2014) are imposed to ensure all text-mining activities happen through the publishers systems.

Contrary to these beliefs, we articulate the need for the development of an open infrastructure offering machine access to data aggregated from many sources, claiming that this machine accessibility layer is necessary for enabling the development of new innovative services. Such machine accessibility layer is not provided by commercial academic search engines, such as Google Scholar, which receive access to research outputs through bespoke arrangements with publishers, but do not technically allow 3rd parties to access this content for text-mining.

The efforts to overcome technical barriers and exploit research outputs have been so far substantially motivated by the worldwide movement towards open science, in particular Open Access. A recent study published by Jisc in June 2014 identifying significant barriers hampering the creation of an *Open Mirror*, which would aggregate all Open Access content, concluded that Jisc should actively seek international support for CORE, covering its full cost in the near term (Jacobs and Ferguson, 2014). This is also motivated by the fact that aggregation systems try to overcome a wide range of technical barriers

to simplify access to the data. Knoth, Rusbridge and Russell (2014) mention primarily the lack of support of the OAI-PMH (Van de Sompel et al., 2004) protocol for content harvesting and the currently insufficient adoption of the ResourceSync protocol by repositories (Klein et al., 2013). This results in the need of using the OAI-PMH protocol in use cases it has never been designed for, making it difficult to decide which data need to be synchronised, identify the licence of the content and to optimise the synchronisation performance. While various guidelines have been issued to increase interoperability, such as the RIOXX Application Profile and OpenAIRE Guidelines (see Section 2.6), these are so far not widely adopted. In spite of these problems, aggregation systems can already provide access to many millions of publications and the amount is steadily increasing as new harvesting approaches are being deployed.

The availability of CORE creates new opportunities for the creation of a wide range of services, many of which make use of text-mining. For example, researchers from Los Alamos National Laboratory use CORE in the context of the *HyberActive* project to mine URLs to associated materials from publication manuscripts for the purposes of archiving them (Shankar et al., 2014). Leetaru et al. (2014) from Georgetown University used CORE as a text-mining dataset for analysing various popular political, social and cultural issues. The OA Button project (Curry, 2014) uses CORE to find Open Access copies of content behind paywalls. Völske et al. (2014) use CORE to automatically in-

duce a classification taxonomy for research papers. A London-based company Researchresearch Ltd. has used CORE data to improve the classification of research funding opportunities and to direct them to appropriate academic communities. The company's CEO William Cullerne Bown said: "As an independent company, we had no obvious access to big, diverse scholarly data - a killer in our drive to develop classification algorithms. The CORE repository, available in bulk, was a breakthrough. Now our algorithms outperform even those from huge publishers." CORE has also become a component of a UNESCOs *Repository for Connecting Local and International Content (CLIC)*. At the World Summit on the Information Society, WSIS-10 Review, Christina von Furstenberg (Senior Programme Specialist, UNESCO) presented the CORE-based CLIC as "the next generation of tools for the [UNESCOs] Management of Social Transformations (MOST) ... that allows comparative access to research, policy recommendations and OA sources, based on semantic analysis." To further improve the awareness about CORE and the possibility of creating new innovative tools using it, we have organised and run so far three International Workshops on Mining Scientific Publications (WOSPs), which were associated with JCDL conferences (Knoth, Zdrahal, Anastasiou, Herrmannova, Jack and Piperidis, 2014; Knoth et al., 2013, 2012).

In October 2011, I started the development of the first CORE prototype, which integrated metadata and content from 61 British repositories and had a

link discovery system that is not significantly different from the existing one. In the meantime I wrote the first application for funding CORE, which I submitted to Jisc. The acceptance of this application allowed me to work on the project for an additional 6 person months. I subsequently wrote more funding proposals, which eventually enabled me to get a small team of developers to help me with the implementation of the CORE system. I am extremely grateful for this support, which made it possible for me to focus on the key aspects of the CORE system development and architecture. I acknowledge the essential help of my colleagues, whose CORE implementation work I supervised at the beginning of the thesis.

8.8 Summary of contribution

The goal of this chapter was to demonstrate how can access to public knowledge be improved through the use of link discovery methods. We argued that a considerable amount of this public knowledge is today present in the form of OA research papers. In order to enable the exploitation of this knowledge, harmonised access to these papers had to be first provided. As such system did not exist, we have decided to develop CORE. We designed CORE based on the concept of three access levels, but stressing, in particular, the importance of programmable access to research outputs. This level of access has not been provided by any system before and CORE now makes the exploitation of this

valuable OA research content possible beyond just link discovery.

Motivated by the ability to apply link discovery methods on the dataset of OA research papers aggregated by CORE, we have discussed how text-based link discovery is applied as part of the overall CORE functionality to detect related as well as duplicate content. We have also demonstrated the ways in which the results produced by link discovery methods can be presented to the user to improve accessibility through the use of plugins and visualisations.

The main original contribution of this chapter is two-fold. Firstly, we have designed the CORE system, as we were motivated by the opportunity to apply link discovery in a domain where exploratory search solutions are needed, but could not be freely built. As of today, CORE has already demonstrated its value and impact in a wide range of use cases going beyond just link discovery. They include primarily text and data mining use cases requiring programmable access to content. CORE should therefore be seen as a direct product of our aim to improve access to public knowledge through pursuing Goal 2 of the thesis. Secondly, we have been the first to apply and provide link discovery on a very large collection of OA research papers, demonstrating the technical feasibility and usefulness of this work in a domain where exploratory search experiences are critically needed.

Chapter 9

Conclusions

Link discovery is an exciting and growing area of research in both academic and corporate environments. It recognises the need to dynamically detect relationships between pieces of information which might otherwise be difficult to spot. There is a wide range of use cases in which link discovery methods can be applied. They include content recommendation, argument analysis, advertising, near-duplicate identification and plagiarism detection (see Chapter 3). By helping to connect pieces of information, we can facilitate the discovery and creation of new knowledge for which the evidence currently remains hidden as it is spread across vast amounts of documents.

In this thesis, we took an approach in which we tried to push the work in this area forward from theory to practice. We started by asking theoretical research questions, such as how people link textual content and how is this related to what can be predicted automatically based on measuring semantic similarity. Answers to these research questions inspired the way we

have thought about semantic similarity in the rest of the thesis. Our findings about the predictive power of semantic similarity discussed in Chapters 4 and 5 have been applied, in particular, in the implementation of the CORE system presented in Chapter 8.

In Chapters 6 and 7, we concentrated on the design, development and evaluation of cross-lingual link discovery systems. We made use of the desirable properties of Wikipedia, originally discussed in Chapter 3, to study the disparity of manually authored links in different languages. This revealed that the link discovery methods presented in Chapter 6 can perform in this task almost as good as humans (for a limited number of recommendations), while being highly scalable. We then shifted our attention, for a moment, to noun phrase-to-document cross-language link discovery use cases. We were inspired by the opportunity to comparatively evaluate our methods with those designed by other research groups in the context of a highly respected international evaluation conference. The evaluations, presented in Chapter 7, showed our methods achieved excellent results in comparison to other participating teams. In NTCIR-10 CrossLink-2, our system dominated the evaluations in the English to Chinese, Japanese and Korean tasks. As we discussed throughout the thesis, comparative evaluation is of key importance for improving link-discovery methods, because the evaluation of systems in domain-specific settings is complicated due to the difficulty of obtaining ground truth judgements. Based on

our experience with the NTCIR CrossLink evaluation, we have identified and described a set of issues in the existing evaluation framework (previously used also at INEX: Link The Wiki Track) and suggested how they can be mitigated or even completely resolved leading to a fairer and more accurate comparison of systems.

Finally, motivated by the opportunity to show how the findings presented in this thesis can be deployed in a real-world application, we have developed a large aggregation system called CORE (Chapter 8). While CORE was designed to support a wide range of use cases requiring access to research outputs, the system has been crucial for demonstrating the need for link discovery in an originally distributed collection of millions of openly accessible documents. CORE also serves as a proof of concept, showing that link discovery techniques can scale up to very large text collections. However, CORE is in the context of the thesis important also for a number of other reasons. Firstly, it shows how programmable access to data and the ability to text-mine the data is crucial for enabling the application of link discovery (and also other types of applications requiring such access). Building new tools or doing research on top of the repository infrastructure used to be very difficult. CORE provides an abstraction harmonising the data coming from hundreds of systems, hence effectively enabling the mining of this content in ways that were not possible before. Secondly, it demonstrates how automatically discovered links might

be presented in a user interface to support exploratory search. Finally, it provides a use case showing that the integration of link discovery techniques with metadata schemas improves metadata consistency, completeness and correctness (Chapter 2) and ensures that the discovered links can also be expressed in a machine readable format and reused by third party applications.

9.1 Summary of contribution

In Section 2.7.2, we identified the research questions and the goals of the thesis. The central research question asked how to support the process of identifying links between semantically related resources and how to facilitate discovery in large textual collections. This question was then, in turn, further broken down into sub-questions we dealt with in the individual chapters of the thesis. In addition, we set ourselves two goals that further contributed to the overall effort. To develop and evaluate new link discovery methods in the context of an international evaluation conference (Chapter 7) and to demonstrate how to facilitate access to public knowledge through the application of link discovery methods (8). We will now provide a short summary of the thesis contributions with respect to the central research question. The detailed summaries of contribution to the research sub-questions (RQ 1-RQ 4) and the research goals (Goal 1 and Goal 2) can be found at the end of Chapters 4-8.

Our first research steps explored the efficiency of identifying links between

textual resources, showing that linking resources can be, with respect to time complexity, seen as the most difficult type of metadata generation (Section 2.3). We explained why dealing with the problem manually is not scalable even in fairly small collections (Table 2.1). Contrary to some existing beliefs (Buckingham Shum and Ferguson, 2010), we argued that automatic methods are necessary as even crowd-sourcing approaches cannot successfully deal with this problem (Chapter 5).

After formalising the link discovery task (Section 3.1), our work continued by reviewing and classifying existing link discovery methods according to granularity (Section 3.2.1), use case (Section 3.2.3) and their input data (Section 3.2.2). According to input data, we divided methods into text-based, semi-structured and link-based approaches. The results of our research (Chapters 6-7) suggest that link-based and semi-structured methods provide better performance than text-based methods, while being typically also less computationally demanding. However, it is important to stress that these methods are largely dependent on collection-specific characteristics and, consequently, it is not possible to generalise across all collections assuming good performance under all conditions. As many (if not the most) textual document collections, such as the collection of research papers aggregated by the CORE system (Chapter 8), are not created with many explicit links forming a densely interconnected network, text-based link discovery methods are of key importance.

One of the main contributions of the thesis in relation to text-based link discovery methods is that we showed semantic similarity can be successfully used as a link predictor, though not exactly in the expected way. This is demonstrated by the negative correlation of semantic similarity and linked-pair likelihood in regions with high similarity (Figure 4.2). We also showed semantic similarity can be used to suggest the link type (Chapter 5). This knowledge has been used in the development of the text-based link discovery methods we applied in the CORE system as part of a content recommendation and duplicates detection system for research papers (Chapter 8).

An important piece of work presented in the thesis is also the development of a number of cross-language link discovery methods (CLLD) for both document-to-document (Chapter 6) and noun phrase-to-document link discovery (Chapter 7). We explained that the multilingual environment creates a very good application domain for the research of link discovery methods. This is due to the ability to study the performance of link discovery methods across languages as well as the potential of applying these methods in a domain where multilingualism causes another significant barrier to the use of manual approaches. We have demonstrated that it is possible to develop effective CLLD methods in these scenarios. In fact, the CLLD methods we submitted to two consecutive evaluation conferences (NTCIR-9 CrossLink and NTCIR-10 CrossLink-2) achieved excellent results in comparison with other

teams in both automatic and manual assessment.

To our knowledge, the methods presented in Chapters 6 and 7, are also the first to utilise ESA and CL-ESA for link discovery. However, perhaps an even more significant contribution was made by documenting the various findings we acquired through the testing and comparison of different CLLD methods. For example, we discovered and discussed the importance of the ranking phase and compared the approach of dealing with the CLLD problem as a similarity search task (Section 7.1) vs a disambiguation task (Section 7.2). This, together with the review of link discovery evaluation approaches (Section 3.3), formed the foundation for being able to explore the limitations of existing link discovery evaluation approaches. One important contribution here was that we showed the low agreement of manually curated link structures created on semantically identical data in different languages. We argued that the use of link structures, created in similar ways, as a ground truth impacts on the measured precision and recall of link discovery systems. This is consistent with the fact that systems in the CrossLink evaluations typically achieve significantly different performance in automatic and manual assessment.

To be able to interpret the results of link discovery systems in some meaningful way, we piloted a new approach to measure the theoretical performance boundary (Section 7.3). This helps us to interpret the performance achieved by link discovery methods and understand the scale of their possible improve-

ment. We were also the first to show how the inter-annotator agreement measured using Cohen's Kappa can be used to link the performance of automatic methods to the performance achieved by people (Section 6.4.3). This experiment showed that at about 5% recall, the inter-annotator agreement of some of our methods is as good as the agreement achieved by independent communities of humans interlinking semantically identical[1] content. While 5% recall might seem as a fairly low recall level, this is, in fact, a very good indicator of performance for content recommendation systems where only the first few recommendations matter.

In order to make the evaluation of link discovery systems more robust, we also proposed a number of improvements. They include the proposition to measure the theoretical performance boundary, change the way results are encoded in CrossLink evaluations, so that methods that rank results according to relevance (instead of confidence) are rewarded, and to use graded relevance in the assessment. However, we also need to stress that the numerical performance of link discovery methods is not the only criterion to be considered in their evaluation. For example, our NTCIR-9 CrossLink methods have been reported to discover qualitatively more novel links rather than just to improve the link graph consistency (Tang, Cavanagh, Trotman, Geva, Xu and Sitbon, 2011). It can be assumed that the choice of input data (links, semi-structured,

[1]The communities interlinked content discussing the same topics, however the descriptions of these topics were only comparable (not completely identical) due to a number of reasons including cultural, language and community size differences.

text) has a significant impact on the qualitative characteristics of the output. Our experience suggests that relying on textual data can lead more likely to the discovery of novel/hidden links, as this approach enables the creation of (unexpected) connections between information with high distance in a manually induced link graph.

Motivated by the possibility of improving the accessibility of information through the automatic discovery of novel (potentially hidden) connections, we realised that the domain of Open Access research papers, which can be seen as a representation of publicly accessible and exploitable knowledge, offers an enormous potential. Our first, yet important contribution to this topic, was the formulation of the need for infrastructures to support the *Three Access Levels* model (Section 8.3), which is needed to exploit this content. While reviewing existing related work (Section 8.2), we showed that such system did not exist and, therefore, we had to create CORE, which has already harvested millions of OA full-text publications from repositories and became a backbone for many projects and research studies.

While the creation of the overall CORE system is certainly a significant contribution of this work, the original use case of CORE was to apply it to support discovery across repositories through the CORE content recommendation plugin. In the thesis, we have shown how the experience we acquired in Chapters 4 and 5 was used to create a monolingual link discovery engine

for CORE to detect related and duplicate content. When faced with the challenge of performing link discovery on a collection of millions of documents, we developed a heuristic method based on a cut-off approach, which makes the solution to scale approximately linearly in time (Section 8.4.3). Although, as we later found out, this solution is similar (but not identical) to the one presented in (Elsayed et al., 2008), the use case and the application domain are new.

We have also discussed how links discovered by CORE can be exposed to users. A new approach described in this thesis is to make use of them as part of a visual search system, which we designed to provide exploratory search in large document collections (Section 8.4.6). The other approaches discussed in the thesis include the use of the CORE recommendation plugin and the CORE Portal. This discovery support has been appreciated by the community of CORE users. For example, Jacobs and Bruce (n.d.) specifically mention that CORE discovery tools are more advanced than those of (a large metadata aggregator) BASE in an article titled *Top ten search engines for researchers that go beyond Google*. Additionally, as we already discussed in the literature review (Chapter 3), automatically generated links can also serve as an input for new applications. Since CORE exposes links through an API and as a downloadable dataset, one such example can be the development of a new class of research impact metrics — *Semantometrics* (Knoth and Her-

rmannova, 2014), which depend on this information. Interestingly, the outputs of these new applications can also provide inputs to further support discovery. For example, impact metrics can be used in search engines to facilitate the discovery of content with high research contribution.

9.2 Limitations

While the thesis focused primarily on performing experiments on openly available datasets (Wikipedia and OA research papers), many of the findings are relevant to any textual collection with similar characteristics. The experiments reported in Chapters 4-7 have been carried out using the latest Wikipedia collection that existed at the time of the experiment. More details on the selection and preprocessing of the datasets is available in the individual chapters. As Wikipedia continually grows and its link structure is being updated, the results achieved by our methods on more recent versions of Wikipedia can slightly differ. However, any such fluctuation is unlikely to impact the fundamental findings of the thesis, such as the relation between semantic similarity and linked-pair likelihood. Please also note that the choice of Wikipedia in our experiments presented in Chapter 5 (intentionally) restricted, as reported, the number of link types. We stress that the value of semantic similarity is therefore only one criterion useful in determining the relationship type. Further research on identifying additional link types, such as contradiction, is needed.

We should also mention that the stability of the presented link discovery methods when applied to different datasets may differ. While the text-based methods can be expected to produce fairly stable results, the performance of link-based approaches is likely to be significantly affected by the specific data collection properties.

Although we have proposed certain changes to the evaluation methodology applied at link discovery evaluation conferences, we do not claim that the existing results reported at these conferences would not be representative of systems' performance. In fact, comparing the automatic and manual assessment results indicates that systems that tend to perform well in the automatic assessment tend to perform well also in the manual assessment. However, as the order of systems according to performance can slightly differ according to the evaluation type, more robust automatic approaches are needed, especially as automatic evaluation is significantly cheaper than manual assessment. To learn more about the statistically significant performance differences of systems presented at CrossLink, we refer the reader to (Tang, Cavanagh, Trotman, Geva, Xu and Sitbon, 2011).

Finally, there is still a range of organisational and technical issues that are limiting the ability of the CORE aggregator to harvest all OA content and perform link discovery over it. These barriers have been documented in (Knoth, Rusbridge and Russell, 2014) and it seems they are fortunately being

slowly removed. Despite these limitations, the dataset with which we worked in Chapter 8 is still sufficiently large to serve as a proof of concept of the scalability and value of applying link discovery in this context.

9.3 Future directions and closing remarks

It can be expected that research in the field of link discovery will continue to thrive and we will see more and more often link discovery solutions being integrated into systems we use on a daily basis. We list below some of the research directions we think will attract further research in the future.

As document-to-document link discovery is becoming more and more prevalent, we can also expect more research to address link discovery where the link source and target are of a lower granularity than a document. Some signs of this are already present on the Internet, such as the interlinking of online newspaper stories, which is similar to the *"wikification"* problem. It can be expected that such technologies will become widely adopted also in other application areas. These will likely include online blogs, various knowledge bases, such as proprietary wikis, and cultural heritage databases, such as online art gallery/museum systems. Similar direction can also be expected in digital libraries, where link discovery has the potential to improve the accessibility of information across online books. When books are long, it is currently especially difficult for people to discover related passages across them. Automatic

methods are capable of assisting in this process already today. In the domain of research papers, it is probably only the matter of time until link discovery (as well as other enrichment technologies based on text-mining) push the industry to move beyond the traditional PDF manuscripts towards richer and more interactive experiences based on linking and comparing information in a document of choice to all other relevant documents on the web.

In order to support the proliferation of applications based on link discovery, a number of challenges still need to be overcome. They relate mainly to the following three areas:

Machine access to information and the ability to text-mine — At the moment, not all web agents are treated equal. Typically, large commercial search engines have much better access to information on the web than new systems. This creates an environment in which only few have access to the data which link discovery technology needs to operate. This is a barrier to the development of new link discovery methods and their application.

Standardisation and harmonisation of relation metadata — A considerable effort is still needed to harmonise the use of metadata on the web which can be used as an input of link discovery techniques and to standardise the outputs which link discovery methods produce, so that applications, such as web browsers, can make effective use of this information.

More sophisticated methods for link typing — The area of automatic link typ-

ing is still not sufficiently researched. New methods for identifying a whole range of link types as well as comparative evaluation studies are needed before these methods can be applied at a scale.

References

About NTCIR (2014), `http://ntcir.nii.ac.jp/about/`. Accessed: 2014-07-08.

Allan, J. (1995), Automatic Hypertext Construction, PhD thesis.

Allan, J. (1996), Automatic Hypertext Link Typing, *in* 'HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext', ACM, New York, NY, USA, pp. 42–52.

Allan, J. (1997), Building Hypertext Using Information Retrieval, Vol. 33, Pergamon Press, Inc., Tarrytown, NY, USA, pp. 145–159.
URL: *http://portal.acm.org/citation.cfm?id=256268.256272*

Almind, T. C. and Ingwersen, P. (1997), 'Informetric analyses on the world wide web: methodological approaches to "webometrics"', *Journal of Documentation* **53**(4), 404–426.

Azzopardi, L. and Vinay, V. (2008), 'Accessibility in Information Retrieval', *Advances in Information Retrieval* pp. 482–489.
URL: *http://link.springer.com/chapter/10.1007/978-3-540-78646-7_46*

Bandura, A., Ross, D. and Ross, S. A. (1961), 'Transmission of aggression through imitation of aggressive models.', *The Journal of Abnormal and Social Psychology* **63**(3), 575.

Berners-Lee, T., Hendler, J., Lassila, O. et al. (2001), 'The Semantic Web', *Scientific American* **284**(5), 28–37.

Bertin, M. and Atanassova, I. (2012), 'Semantic Enrichment of Scientific Publications and Metadata: Citation Analysis Through Contextual and Cognitive Analysis', *D-Lib Magazine* **18**(7/8).

Bible Data Files (2014), `http://www.sacred-texts.com/bib/osrc/index.htm`. Accessed: 2014-07-07.

Bizer, C., Heath, T. and Berners-Lee, T. (2009), 'Linked Data - The Story So Far', *International Journal on Semantic Web and Information Systems (IJSWIS)* **5**(3), 1–22.

URL: *http://dx.doi.org/10.4018/jswis.2009081901*

Björk, B.-C. (2003), 'Open Access to Scientific Publications - An Analysis of the Barriers to Change', *Inf. Res.* **9**(2).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* **3**, 993–1022.

Blustein, W. J. (1999), Hypertext Versions of Journal Articles: Computer Aided linking and realistic human-based evaluation, PhD thesis.

Bruner, J. S. (1961), 'The act of discovery.', *Harvard educational review* .

Buckingham Shum, S. and De Liddo, A. (2010), 'Collective Intelligence for OER Sustainability'.

Buckingham Shum, S. and Ferguson, R. (2010), Towards a Social Learning Space for Open Educational Resources, *in* 'OpenED2010: Seventh Annual Open Education Conference'.
URL: *http://oro.open.ac.uk/23351/*

Buckingham Shum, S., Motta, E. and Domingue, J. (2000), 'ScholOnto: an Ontology-Based Digital Library Server for Research Documents and Discourse', *International Journal on Digital Libraries* **3**, 237–248.
URL: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.3835*

Bush, V. (1945), 'As We May Think', *The Atlantic Monthly* .

Caragea, C., Silvescu, A., Mitra, P. and Giles, C. L. (2013), Can'T See the Forest for the Trees?: A Citation Recommendation System, *in* 'Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries', JCDL '13, ACM, New York, NY, USA, pp. 111–114.
URL: *http://doi.acm.org/10.1145/2467696.2467743*

Cesa-Bianchi, N., Gentile, C. and Zaniboni, L. (2006), Hierarchical Classification: Combining Bayes with SVM, *in* 'ICML '06: Proceedings of the 23rd International Conference on Machine Learning', ACM, New York, NY, USA, pp. 177–184.

Charikar, M. S. (2002), Similarity Estimation Techniques from Rounding Algorithms, *in* 'Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing', STOC '02, ACM, New York, NY, USA, pp. 380–388. **URL:** *http://doi.acm.org/10.1145/509907.509965*

Chen, F., Farahat, A. and Brants, T. (2004), Multiple Similarity Measures and Source-Pair Information in Story Link Detection, *in* 'HLT-NAACL', pp. 313–320.

Cimiano, P. (2006), *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Curry, S. (2014), 'Push Button for Open Access', The Guardian. **URL:** *http://www.theguardian.com/science/2013/nov/18/open-access-button-push*

Daniel, R. (2012), 'Domain-Independent Mining of Abstracts Using Indicator Phrases', *D-Lib Magazine* **18**(7/8).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990), Indexing by latent semantic analysis, Vol. 41, pp. 391–407.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990), 'Indexing by Latent Semantic Analysis', *Journal of the American Society for Information Science* **41**, 391–407.

Dekel, O., Keshet, J. and Singer, Y. (2004), Large margin hierarchical classification, *in* 'ICML '04: Proceedings of the twenty-first international conference on Machine learning', ACM, New York, NY, USA, p. 27.

Ellis, D., Furner-Hines, J. and Willett, P. (1994), On the Measurement of Inter-Linker Consistency and Retrieval Effectiveness in Hypertext Databases, *in* 'SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval', Springer-Verlag New York, Inc., New York, NY, USA, pp. 51–60.

Elsayed, T., Lin, J. and Oard, D. W. (2008), 'Pairwise Document Similarity in Large Collections with MapReduce', (June), 265–268.

Erbs, N., Gurevych, I. and Rittberger, M. (2013), 'Bringing Order to Digital Libraries: From Keyphrase Extraction to Index Term Assignment', *D-Lib Magazine* **19**(9/10).

Erbs, N., Zesch, T. and Gurevych, I. (2011), Link Discovery: A Comprehensive Analysis, *in* 'Proceedings of the 5th IEEE International Conference on Semantic Computing (IEEE-ICSC)', p. to appear.

Fahrni, A., Nastase, V. and Strube, M. (2011), 'HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task', *Proceedings of NTCIR-9* pp. 473–480.

Frommholz, I. (2001), Categorizing web documents in hierarchical catalogues, *in* 'In Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research'.

Gabrilovich, E. and Markovitch, S. (2007), Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, *in* 'In Proceedings of the 20th International Joint Conference on Artificial Intelligence', pp. 1606–1611.

Garfield, E. et al. (1965), Can Citation Indexing be Automated, *in* 'Statistical association methods for mechanized documentation, symposium proceedings', pp. 189–192.

Geva, S. (2007), GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia, *in* N. Fuhr, J. Kamps, M. Lalmas and A. Trotman, eds, 'INEX', Lecture Notes in Computer Science, Springer.

Geva, S., Kamps, J. and Trotman, A., eds (2009), *Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers*, Lecture Notes in Computer Science, Springer.

Granitzer, M., Seifert, C. and Zechner, M. (2008), Context Based Wikipedia Linking, *in* Geva et al. (2009), pp. 354–365.

Green, S. J. (1998), 'Automated link generation: can we do better than term repetition?', *Comput. Netw. ISDN Syst.* **30**(1-7), 75–84.

Green, S. J. (1999), 'Building Hypertext Links By Computing Semantic Similarity', *IEEE Trans. on Knowl. and Data Eng.* **11**(5), 713–730.

Greenberg, J. (2004), 'Metadata Extraction and Harvesting', **6**(4), 59–82.

Greenberg, J., Spurgin, K. and Crystal, A. (2006), 'Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions', *International Journal of Metadata, Semantics and Ontologies* **1**(1), 3.
URL: *http://www.inderscience.com/link.php?id=8766*

Hansen, W. G. (1959), 'How Accessibility Shapes Land Use', *Journal of the American Institute of Planners* **25**(2), 73–76.
URL: *http://dx.doi.org/10.1080/01944365908978307*

Herrmannova, D. and Knoth, P. (2012), 'Visual Search for Supporting Content Exploration in Large Document Collections', *D-Lib Magazine* **18**(7/8).

Hoffart, J., Zesch, T. and Gurevych, I. (2009), 'An Architecture to Support Intelligent User Interfaces for Wikis by Means of Natural Language Processing Categories and Subject Descriptors Wiki-Mining'.

Huang, D. W., Xu, Y., Trotman, A. and Geva, S. (2008), Focused Access to XML Documents, Springer-Verlag, Berlin, Heidelberg, chapter Overview of INEX 2007 Link the Wiki Track, pp. 373–387.

Huang, W. C., Geva, S. and Trotman, A. (2009), 'Overview of the INEX 2009 Link the Wiki Track'.

Huang, W. C., Trotman, A. and Geva, S. (2008), 'Experiments and evaluation of link discovery in the wikipedia'.

Ikeda, D. (2006), 'Automatically Linking News articles to Blog entries'.

Implementing the Hargreaves review (2014), `http://www.ipo.gov.uk/types/hargreaves.htm`. Accessed: 2014-07-07.

Initiative, O. A. (2002), 'Protocol for metadata harvesting, v2.0'.
URL: *http://www.openarchives.org/OAI/openarchivesprotocol.html*

Itakura, K. Y. and Clarke, C. L. A. (2008), University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks, *in* Geva et al. (2009), pp. 132–139.

Jacobs, N. and Bruce, R. (n.d.), 'Ten Search Engines for Researchers that Go Beyond Google, url = http://www.jisc.ac.uk/inform/inform37/SearchingBeyondGoogle.html, note = Accessed: 2014-06-03'.

Jacobs, N. and Ferguson, N. (2014), 'Bringing the UK's open access research outputs together: barriers on the Berlin road to open access', Jisc report.
URL: *http://www.jisc.ac.uk/publications/reports/2014/bringing-the-uks-open-access-research-outputs-together.aspx*

Jacobson, K., Raimond, Y. and Gangler, T. (2010), 'The similarity ontology - musim'.
URL: *http://kakapo.dcs.qmul.ac.uk/ontology/musim/0.2/musim.html*

Järvelin, K. and Kekäläinen, J. (2002), 'Cumulated Gain-based Evaluation of IR Techniques', *ACM Trans. Inf. Syst.* **20**(4), 422–446.
URL: *http://doi.acm.org/10.1145/582415.582418*

Jenkinson, D., Leung, K.-C. and Trotman, A. (2008), Wikisearching and Wikilinking, *in* Geva et al. (2009), pp. 374–388.

Jiang, Zhuoren and Liu, X. (2013), 'Recovering Missing Citations in a Schholarly Network', pp. 419–420.

Joachims, T. (2006), Training Linear SVMs in Linear Time, *in* 'Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery

and data mining', KDD '06, ACM, New York, NY, USA, pp. 217–226.
URL: *http://doi.acm.org/10.1145/1150402.1150429*

Kamps, J., Geva, S. and Trotman, A. (2010), Analysis of the inex 2009 ad hoc track results, *in* 'Focused retrieval and evaluation', Springer, pp. 26–48.

Katukuri, J., Mukherjee, R. and Konik, T. (2013), 'Large-scale Recommendations in a Dynamic Marketplace'.

Kern, R., Jack, K., Hristakeva, M. and Granitzer, M. (2012), 'TeamBeam - Meta-Data Extraction from Scientific Literature', *D-Lib Magazine* **18**(7/8).

Kittur, A., Suh, B., Pendleton, B. A. and Chi, E. H. (2007), He Says, She Says: Conflict and Coordination in Wikipedia, *in* 'CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, New York, NY, USA, pp. 453–462.
URL: *http://dx.doi.org/10.1145/1240624.1240698*

Klein, M., Sanderson, R., Van de Sompel, H., Warner, S., Haslhofer, B., Lagoze, C. and Nelson, M. L. (2013), 'A technical framework for resource synchronization', *D-Lib Magazine* **19**(1), 3.

Knoth, P. (2013), 'From open access metadata to open access content: Two principles for increased visibility of open access content'.

Knoth, P. (2014), 'The DiggiCORE project: Aggregating and mining the world of open access articles - draft'.

Knoth, P., Anastasiou, L. and Pearce, S. (2014), 'My repository is being aggregated: a blessing or a curse?'.

Knoth, P. and Herrmannova, D. (2013), 'Simple yet effective methods for cross-lingual link discovery (CLLD)-KMI@ NTCIR-10 CrossLink-2'.

Knoth, P. and Herrmannova, D. (2014), 'Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing Research Contribution'.

Knoth, P., Rusbridge, A. and Russell, R. (2014), 'Open Mirror Feasibility Study, Appendix A: Technical Prototyping Report', Jisc report.
URL: *http://www.jisc.ac.uk/media/documents/publications/reports/2014/OpenMirror_Full_*

Knoth, P., Schmidt, M., Smrz, P. and Zdrahal, Z. (2009), 'Towards a Framework for Comparing Automatic Term Recognition Methods'.

Knoth, P. and Zdrahal, Z. (2011), Mining Cross-document Relationships from Text, *in* 'The First International Conference on Advances in Information Mining and Management (IMMM 2011)'.

Knoth, P., Zdrahal, Z., Anastasiou, L., Herrmannova, D., Jack, K. and Piperidis, S. (2014), 'Guest Editorial: Special Issue on Mining Scientific Publications', *To appear in D-Lib Magazine* **20**(10/11).

Knoth, P., Zdrahal, Z., Freire, N. and Muhr, M. (2013), 'Scientific Publications: Gathering Data, Extracting Information, and Following Trends', *D-Lib Magazine* **19**(9/10).

Knoth, P., Zdrahal, Z. and Juffinger, A. (2012), 'Guest Editorial: Special Issue on Mining Scientific Publications', *D-Lib Magazine* **18**(7/8).

Knoth, P., Zilka, L. and Zdrahal, Z. (2011), KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis, *in* 'NTCIR-9'.

Kolak, O. and Schilit, B. N. (2008), Generating links by mining quotations, *in* 'HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia', ACM, New York, NY, USA, pp. 117–126.

Kopetz, H. (2011), Internet of Things, *in* 'Real-time systems', Springer, pp. 307–323.

Lagoze, C., Krafft, D. B., Payette, S. and Jesuroga, S. (2005), 'What is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL', *D-Lib Magazine* **11**(11).

Leetaru, K. H., Perkins, T. K. and Rewerts, C. (2014), 'Cultural Computing at Literature Scale', *D-Lib Magazine* **20**(9/10).

Liu, T.-Y. (2009), 'Learning to Rank for Information Retrieval', *Found. Trends*

Inf. Retr. **3**(3), 225–331.

URL: *http://dx.doi.org/10.1561/1500000016*

Loesch, M. F. (2010), 'OAIster Database http://oaister.worldcat.org/', *Technical Services Quarterly* **27**(4), 395–396.

URL: *http://www.tandfonline.com/doi/abs/10.1080/07317131.2010.501001*

Lu, W., Liu, D. and Fu, Z. (2008), CSIR at INEX 2008 Link-the-Wiki Track, *in* Geva et al. (2009), pp. 389–394.

Lyte, V., Jones, S., Ananiadou, S. and Kerr, L. (2009), 'UK Institutional Repository Search: Innovation and Discovery', *Ariadne* **61**.

URL: *http://www.ariadne.ac.uk/issue61/lyte-et-al/*

Mancini, C. (2005), *Cinematic Hypertext: Investigating a New Paradigm.*

Manghi, P., Mikulicic, M., Candela, L., Artini, M. and Bardi, A. (2010), General-Purpose Digital Library Content Laboratory Systems, *in* 'ECDL', pp. 14–21.

Manghi, P., Mikulicic, M., Candela, L., Castelli, D. and Pagano, P. (2010), 'Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System', *D-Lib Magazine* **16**(3/4).

Manku, G. S. (2007), 'Detecting Near-Duplicates for Web Crawling', pp. 141–149.

Manning, C. D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge.

URL: *http://www-csli.stanford.edu/ hinrich/information-retrieval-book.html*

Manning, C. D. and Schuetze, H. (1999), *Foundations of Statistical Natural Language Processing*, 1 edn, The MIT Press.

URL: *http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262133601*

Marchionini, G. (2006), 'Exploratory Search: from Finding to Understanding', *Commun. ACM* **49**(4), 41–46.

URL: *http://doi.acm.org/10.1145/1121949.1121979*

McDonald, D. and Kelly, U. (2012), Value and Benefits of Text Mining, *in* 'JISC Report'.

McNamee, P., Hltcoe, J. H. U. and Ldc, H. S. (2009), 'Overview of the TAC 2009 Knowledge Base Population Track', (November).

Mihalcea, R. and Csomai, A. (2007), Wikify!: Linking Documents to Encyclopedic Knowledge, *in* 'Proceedings of the sixteenth ACM conference on Conference on information and knowledge management', CIKM '07, ACM, New York, NY, USA, pp. 233–242.

URL: *http://doi.acm.org/10.1145/1321440.1321475*

Milne, D. and Witten, I. H. (2007), 'An Effective , Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links', pp. 25–30.

Milne, D. and Witten, I. H. (2008), Learning to Link with Wikipedia, *in* 'Proceeding of the 17th ACM conference on Information and knowledge management', CIKM '08, ACM, New York, NY, USA, pp. 509–518.
URL: *http://doi.acm.org/10.1145/1458082.1458150*

Petrič, I., Urbančič, T., Cestnik, B. and Macedoni-Lukšič, M. (2009), 'Literature mining method rajolink for uncovering relations between biomedical concepts', *Journal of Biomedical Informatics* **42**(2), 219 – 227.
URL: *http://www.sciencedirect.com/science/article/pii/S1532046408001044*

Pieper, D. and Summann, F. (2006), 'Bielefeld Academic Search Engine (BASE): an end-user oriented institutional repository search service', *Library Hi Tech* **24**(4), 614 – 619.

Potthast, M., Gollub, T. and Hagen, M. (2012), 'Overview of the 4th International Competition on Plagiarism Detection.', *... Online Working Notes ...* pp. 23–26.
URL: *http://users.dsic.upv.es/ prosso/resources/PotthastEtAl_PAN_CLEF13.pdf*

Priem, J., Taraborelli, D., Groth, P. and Neylon, C. (2010), 'Altmetrics: A Manifesto'.

Rabin, M. (1981), Fingerprinting by Random Polynomials, Technical report, Center of Research in Computer Technology.

Radev, D. R., Zhang, Z. and Otterbacher, J. (2008), 'Cross-Document Relationship Classification for Text Summarization', *Unpublished paper* .

Reilly, S. (2014), Realising the Innovative Potential of Digital Research Methods: A Call from the Research Community, *in* 'LIBER'.
URL: *http://libereurope.eu/wp-content/uploads/2014/07/Open-Letter-To-Elsevier1.pdf*

Reynar, J. (1998), Topic Segmentation: Algorithms and Applications, PhD thesis.

Robinson, P. (2014), 'Writing a dissertation', `http://www.cl.cam.ac.uk/~pr10/teaching/dissertation.html`. Accessed: 2014-23-09.

Rodrigues, E. and Clobridge, A. (2011), 'The Case for Interoperability for Open Access Repositories', *Working Group 2: Repository Interoperability* .
URL: *http://www.coar-repositories.org/files/A-Case-for-Interoperability-Final-Version.pdf*

Rose, D. E. and Levinson, D. (2004), Understanding User Goals in Web Search, *in* 'Proceedings of the 13th international conference on World Wide Web', WWW '04, ACM, New York, NY, USA, pp. 13–19.
URL: *http://doi.acm.org/10.1145/988672.988675*

Sakai, T. (2009), 'On the Robustness of Information Retrieval Metrics to Biased Relevance Assessments', *JIP* **17**, 156–166.

Sebastiani, F. (2002), 'Machine Learning in Automated Text Categorization', *ACM Computing Surveys* **34**(1), 1–47.

URL: *http://citeseer.ist.psu.edu/518620.html*

Shankar, H., Klein, M. and Sompel, H. V. d. (2014), 'HyberActive Demo', Demo.

URL: *https://www.dropbox.com/s/rysi5ekmozbipyq/4min_video.mp4*

Shotton, D. (2010), 'CiTO, the Citation Typing Ontology.', *Journal of biomedical semantics* **1 Suppl 1**(Suppl 1), S6.

URL: *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2903725&tool=pmcentrez&rendertype=abstract*

Smalheiser, N. R. and Swanson, D. R. (1998), 'Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses', *Computer Methods and Programs in Biomedicine* **57**(3), 149 – 153.

URL: *http://www.sciencedirect.com/science/article/pii/S0169260798000339*

Smet, W. D. (2009), 'Cross-Language Linking of News Stories on the Web using Interlingual Topic Modelling'.

Sorg, P. and Cimiano, P. (2008*a*), Cross-lingual Information Retrieval with

Explicit Semantic Analysis, *in* 'Working Notes for the CLEF 2008 Workshop'.

Sorg, P. and Cimiano, P. (2008*b*), Enriching the Crosslingual Link Structure of Wikipedia - A Classification-Based Approach -, *in* 'Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WikiAI'08)', To appear.

Strohman, T., Croft, W. B. and Jensen, D. (2007), 'Recommending Citations for Academic Papers', *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* p. 705.
URL: *http://portal.acm.org/citation.cfm?doid=1277741.1277868*

Swanson, D. R. (1986), 'Undiscovered Public Knowledge', *The Library Quarterly: Information, Community, Policy* **56**(2), 103 – 118.
URL: *http://www.sciencedirect.com/science/article/pii/S0169260798000339*

Tang, L.-X., Cavanagh, D., Trotman, A., Geva, S., Xu, Y. and Sitbon, L. (2011), 'Automated Cross-Lingual Link Discovery in Wikipedia', pp. 512–519.

Tang, L.-X., Geva, S., Trotman, A., Xu, Y. and Itakura, K. (2011), Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery, *in* 'NTCIR-9'.

Tang, L.-X., Geva, S., Trotman, A., Xu, Y. and Itakura, K. Y. (2014), 'An Evaluation Framework for Cross-Lingual Link Discovery', *Information Processing & Management* **50**(1), 1 – 23.

URL: *http://www.sciencedirect.com/science/article/pii/S0306457313000757*

Tang, L.-X., Itakura, K., Geva, S., Trotman, A. and Xu, Y. (2011), The Effectiveness of Cross-Lingual Link Discovery, *in* 'Proceedings of The Fourth International Workshop on Evaluating Information Access (EVIA)', pp. 1–8.

Tang, L.-X., Kang, I.-S., Kimura, F., Lee, Y.-H., Trotman, A., Geva, S. and Xu, Y. (2013), Overview of the NTCIR-10 Cross-Lingual Link Discovery Task, *in* 'NTCIR-10'.

Teufel, S., Siddharthan, A. and Tidhar, D. (2006), 'Automatic Classification of Citation Function', *... of the 2006 Conference on Empirical ...* .

URL: *http://dl.acm.org/citation.cfm?id=1610091*

The Bible Tops the List of the Most Read Books in the World (2014), `http://www.relevantmagazine.com/slices/bible-tops-list-most-read-books-world`. Accessed: 2014-07-07.

Trigg, R. (1983), A Network-Based Approach to Text Handling for the Online Scientific Community, PhD thesis.

Trotman, A., Alexander, D. and Geva, S. (2009), 'Overview of the INEX 2010 Link the Wiki Track'.

Tsagkias, M., Rijke, M. D. and Weerkamp, W. (2011), 'Linking Online News and Social Media'.

Uren, V., Buckingham Shum, S., Li, G., Domingue, J. and Motta, E. (2003), Scholarly Publishing and Argument in Hyperspace, *in* 'Proceedings of the 12th International Conference on World Wide Web', WWW '03, ACM, New York, NY, USA, pp. 244–250.

URL: *http://doi.acm.org/10.1145/775152.775187*

Use of I, we and the Passive Voice in a Scientific Thesis (2014), `http://english.stackexchange.com/questions/24629/use-of-i-we-and-the-passive-voice-in-a-scientific-thesis`. Accessed: 2014-23-09.

Van de Sompel, H., Nelson, M. L., Lagoze, C. and Warner, S. (2004), 'Resource harvesting within the oai-pmh framework', *D-lib magazine* **10**(12), 1082–9873.

Völske, M., Gollub, T., Hagen, M. and Stein, B. (2014), 'Keyquery-Based Classification System for CORE', *D-Lib Magazine* **20**(10/11).

Volz, J., Bizer, C., Gaedke, M. and Kobilarov, G. (2009), 'Silk-A Link Discovery Framework for the Web of Data', *LDOW* **538**.

Weeber, M., Klein, H., de Jong-van den Berg, L. T. and Vos, R. (2001), 'Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-

Fish Oil and Migraine-Magnesium discoveries', *Journal of the American Society for Information Science and Technology* **52**(7), 548–557.

URL: *http://dx.doi.org/10.1002/asi.1104*

White, B., Jackman, A., Marcich, C., Mollet, R., Ellis, P., Wishart, P., McVay, J., Killock, J. and Ashcroft, R. (2011), 'Corrected Transcript of Oral Evidence', Statement to Houses of Parliament - Q 187.

URL: *http://www.publications.parliament.uk/pa/cm201012/cmselect/cmbis/c1498-iii/c149801.htm*

Widdows, D. and Ferraro, K. (2008), Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application, *in* B. M. J. M. J. O. S. P. D. T. Nicoletta Calzolari (Conference Chair), Khalid Choukri, ed., 'Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)', European Language Resources Association (ELRA), Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/.

Wilkinson, R. and Smeaton, A. F. (1999), 'Automatic Link Generation', *ACM Computing Surveys* **31**.

Yang, H. and Callan, J. (2006), Near-duplicate Detection by Instance-level Constrained Clustering, *in* 'Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Re-

trieval', SIGIR '06, ACM, New York, NY, USA, pp. 421–428.

URL: *http://doi.acm.org/10.1145/1148170.1148243*

Zeng, J. and Bloniarz, P. A. (2004), 'From Keywords to Links: an Automatic Approach', *Information Technology: Coding and Computing, International Conference on* **1**, 283.

Zhang, J. and Kamps, J. (2009), 'A Content-Based Link Detection Approach', pp. 395–400.