# Mining Cross-document Relationships from Text

Petr Knoth
*Knowledge Media Institute*
*The Open University*
*Milton Keynes, United Kingdom*
*p.knoth@open.ac.uk*

Zdenek Zdrahal
*Knowledge Media Institute*
*The Open University*
*Milton Keynes, United Kingdom*
*z.zdrahal@open.ac.uk*

***Abstract*—The paper argues that automatic link generation and typing methods are needed to find and maintain cross-document links in large and growing textual collections. Such links are important to organise information and to support search and navigation. We present an experimental study on mining cross-document links from a collection of $5000$ documents. We identify a set of link types and show that the value of semantic similarity is a good distinguishing indicator.**

***Keywords*-text mining, automatic link generation and typing, semantic similarity, digital libraries**

## I. Introduction

There has been a significant research effort in the area of modelling cross-document relationships. These include various 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 inter-document relationships provides the means for better organising fragmented information. In practise, this would improve the browsing, the navigation and the discovery of important information resources. However, the current cross-document relationship modelling approaches rely on human annotators and therefore they do not scale-up. So far, little work has seriously addressed the limitations of manual identification of cross-document relationships in large and constantly growing repositories. In this paper, we argue that automatic link discovery and typing methods[1] can be used to bridge this gap.

This work is based on the following hypotheses:

(a) Cross-document links can be generated automatically using semantic similarity as one of the criteria.
(b) The value of semantic similarity is related to the link type.

The paper brings the following contributions:

- We provide evidence and argue why automatic link generation is necessary for the creation and maintenance of typed relationships, especially in scholarly databases and encyclopedias, and why it cannot be easily substituted by social tagging or crowdsourcing approaches.
- We elaborate the abovementioned hypotheses, especially (b), for a selection of link types.
- We present a simple experiment for mining link types motivated by the results previously reported in [1].

The rest of the paper is organized as follows. In Section II, the role automatic link generation methods can play in automatically analyzing large text collections is introduced. Related work in the areas of semantic web tools for discourse modeling, automatic link generation and link typing is discussed in Section III. In Section IV, we argue why automatic link generation is needed and cannot be substituted by crowdsourcing approaches. An experimental link typing study is presented in Section V. Finally, the paper is concluded in Section VI.

## II. Automatic mining of cross-document links

The automatic link generation task can be defined as follows: Let $S$ and $T$ be collections of documents, denoting sources and targets respectively. Let $s \in S$ and $t \in T$ be lexical units of possibly different granularity. For example, $s$ and $t$ can be the whole documents, paragraphs, sentences or even noun phrases. The goal is to find a binary relation $\rho \subseteq S \times T$ defined in terms of pairs $< s_i, t_j >$ such that all pairs are interpreted by a human evaluator as carrying the same semantic relationship. For example, $\rho$ can be interpreted as *is similar*, *is_the_same*, *expands*, *contradicts* etc. The relation must satisfy the usual properties, e.g. *is_the_same* is symmetric, transitive and reflexive, *is_similar* is not transitive, *expands* is antisymmetric etc.

Automatic link generation methods have many potential applications. For example, the methods can be used for the interlinking of resources not originally created as hypertext documents, for the maintenance or the discovery of new links in collections growing in size, or to improve navigation in collections with long texts, such as books or newspaper articles. All this makes the automatic mining of cross-document links a very useful technology which could be applied across a number of disciplines including information retrieval, semantic web, user navigation, text summarization and others.

[1] In this paper, the concept of *link* refers to a semantic connection between two segments of text, such as two documents or paragraphs, at the discourse level and should not be confused with the Semantic Web representation known as Open Linked Data, which is an approach for publishing data and their relations using RDF triples.

## III. Related work

### A. Semantic web technology for cross-document relationship modeling

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. These systems currently do not provide support for organising, modeling and sharing cross-document relationships. Consequently, their navigation capabilities are limited.

To fill the gap, scientific community invested significant effort into relationship and argument modeling tools. For example, the Mendeley tool [2] allows to discover related research literature, highlight and organise it, annotate relationships to other articles and share them with others. Similar work has been done previously by Uren et. al., the ClaiMaker tool described in [3] allows to model and share research debates/discourses across scientific literature. Other work has also focused on relationship and argument visualization [4]. A number of tools have also been developed for specific domains, such as the life sciences [5], [6].

Though the abovementioned studies recognise the potential offered by collaborative tagging, crowdsourcing and sharing, the resulting approaches rely in the end always on human annotators. We claim that there are at least two reasons why such an approach cannot scale-up: (1) The rate of information growth is faster than the resources of the crowds. This issue is further discussed in Section IV. (2) Researchers are usually reluctant to share this type of knowledge, because the skill of analyzing and interpreting papers is the researcher's know-how. This has also been recognised in the tool presentd in [3] where sharing is restricted to a selected research community.

### B. Link generation

In the 1990s, the main application area for link generation methods were hypertext construction systems [7]. Nowadays, link generation methods for finding related documents have become the de-facto standard. They have been applied in large digital repositories, such as PubMed or the ACM Digital Library, or in search engines including Google Scholar. Generating links pointing to units of a lower granularity than a document has been investigated more recently. The task of such systems is to locate relevant information inside the document instead of only providing a link to the whole document. The Initiative for the Evaluation of XML retrieval (INEX) played an important role in the link generation research by providing evaluation tracks (Link-the-Wiki track) for link generation systems at the granularity of documents as well as at a more fine-grained granularity [8].

Current approaches can be divided into three groups: (1) *link-based* approaches discover new links by exploiting an existing link graph [9], [10], [11]. (2) *semi-structured* approaches try to discover new links using semi-structured information, such as the anchor texts or document titles [12], [13], [14]. (3) *purely content-based* approaches use as an input plain text only. They typically discover related resources by calculating semantic similarity based on document vectors [15], [16], [17], [18]. Some of the mentioned approaches, such as [11], combine multiple methods.

### C. Link taxonomies/ontologies and link typing

A pioneering study in link typing has been presented already in 1980s by Randall Trigg [19] who developed a taxonomy of link types. Trigg divided links into two groups - normal (inter-document) links and commentary (cross-document) links. His rich taxonomy of link types enables the specification of judgements on hypertext nodes. With link types, such as *unimportant*, *solved*, *insufficient* or *incoherent* the taxonomy is *content focused* rather than *relation focused* [20]. Another approach is represented by the ScholOnto taxonomy [21] which has been developed with a reference to cognitive coherence relations [22].

An influential study on automatic generation and typing of links has been published in [23]. Allan recognizes that certain cross-document link types (*automatic links*) can be automatically extracted more easily than others. He focuses then on the development of methods for the identification of the automatic link types, involving relations such as *tangent*, *equivalence* or *contrast*.

An unsupervised approach for the recognition of discourse relations has been presented in [24]. The authors show that from a set of adjacent sentences a subset of discourse relations, namely *contrast*, *explanation-evidence*, *condition* and *elaboration* can be recognized with high accuracy. This task is significantly more difficult, but also more interesting, in the cross-document settings. Similar problem has been recently addressed in Radev et. al. [25] who introduced a taxonomy of 18 cross-document rhetorical relationships denoted as Cross-document Structure Theory (CST). In addition, they present the development of an annotated dataset of CST relationships and experiment with the recognition of their subset using machine learning with a varying level of success for different relationships.

## IV. Manual annotation and crowdsourcing

Cross-document discourse modeling, i.e. connecting a claim found in one document with a claim found in another one by a semantic relation, such as *contradicts*, is technically identical to the problem of providing metadata that allow to organize resources and information in a logical way. Various social annotation tools for metadata generation available on the Web have become very popular, such as image tagging or rating systems. Most applications that use them are based on the idea that a large number of users can provide in most cases good quality metadata. However, there is a number

of problems where the knowledge of the crowds is not sufficient due to lack of human expertise or theoretical time constraints.

It has been shown [26] that metadata can be divided into three distinct groups with respect to the nature of information they are describing. (1) Metadata describing the content of a resource (2) Metadata classifying a resource using a taxonomy (3) Metadata connecting two resources usually by a semantic relation. While provision of type (1) metadata can be done by humans for large text collections in a reasonable time, the provision of type (2) metadata is problematic and type (3) metadata cannot be manually acquired even in moderately large collections. The reason is that the number of possible connections explodes quadratically with respect to the number of resources and as a result people are unable to keep track of all the relevant available information. The problem appears to be particularly significant in quickly growing collections with many contributing authors.

A tempting approach to resolve this problem is by increasing the number of people who contribute to the collection maintenance, for example, by creating discourse links and then sharing the results with others. Shum and Fergusson expect that this 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 [27]. In reality, this approach can be successful only in very limited domains, it certainly does not scale-up unless automatic link generation and typing tools assist in the annotation and the maintenance process. In addition to that, human annotators have been previously found inconsistent in carrying out this task [28].

To provide an example, let us consider Wikipedia, which is today perhaps the largest collection of documents containing user created links and at the same time maintained by a very large community of users (about 250,000 contributing users). Even though Wikipedia contains currently 3,433,587 articles, it is still very small in comparison to all information available on the Web. While in Wikipedia content is typically linked from an anchor (concept) to the whole article (description of the concept), the situation is more complex in other domains, such as in scholarly databases. In Wikipedia, there can be only one page describing a concept whereas in scholarly databases there can be a large and growing number of papers discussing the same topic. The growth of Wikipedia in terms of new articles has already started to decrease and it is predicted that this trend is going to continue in the future. An opposite trend can be expected with scholarly literature.

Even though the problem of linking information less complex in Wikipedia than in scholarly databases and even though the community is very large, the maintenance of Wikipedia is problematic and automatic tools are desperately needed. For example, it has been noted in [29] that the effort necessary for the maintenance of the information on Wikipedia is not directly proportional to the amount of information stored, but rises faster than linearly with the amount of information being added.

## V. Using semantic similarity for link typing

We have previously studied the relation between links authored by people and links predicted by automatic link generation methods [1], namely using semantic similarity measures on document vectors directly extracted from text. The results indicate that semantic similarity is strongly correlated to the way people link content. In this paper, we are extending this work by investigating the qualitative properties of links. As a test-bed we are using articles selected from Wikipedia. For our experiments, this dataset has the following advantages:

- A large number of good quality articles forming a network of cross-references created and agreed by a sufficiently large community of Wikipedia contributors.
- Articles connected by a single *unspecified* link type. However, the link may represent different semantic relationships.
- A suitable initial test-bed. 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 that typically do not appear at this level.

The correlation has been measured on a collection of 5,000 Wikipedia articles in categories containing the phrase "United Kingdom". This required the calculation of semantic similarity (in this case *cosine similarity* calculated on *tfidf* document vectors) for $\frac{5{,}000^2}{2} - 5{,}000 = 12{,}495{,}000$ pairs of documents and the extraction of all 120,602 links between these articles created by Wikipedia authors.

### A. Linked-pair likelihood

A central concept of our study is the quantity called *linked-pair likelihood* introduced in [1] 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 1 shows $lpr$ calculated for groups of document pairs at different intervals of semantic similarity. It can be observed that linked-pair likelihood strongly correlates with the value of sematic similarity (this provides an answer to hypothesis (a) in the introduction), however the direction of the correlation is in the right part of the graph quite unexpected. The correlation has been tested for statistical significance with a positive result for $p$-value well beyond $p < 0.001$ for both Spearman's rank and Pearson correlation coefficients. This indicates that high similarity value is not necessarily a good predictor for the existence of a link. The detail of this experiement can be found in [1].

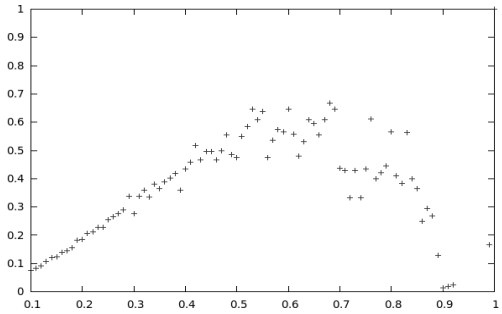

Figure 1. The linked-pair likelihood (y-axis) with respect to the cosine similarity (x-axis) [1].

### B. Using semantic similarity for relation typing

The results presented in the previous section provoke a number of questions. Perhaps the two most interesting are:

(1) Why is the curve in Figure 1 not monotonically increasing which would mean the more semantic similar the more likely 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 hypothesize 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. Investigation of these two questions provides answers to hypothesis (b) presented in the introduction.

### C. Relations of interest and their representation

In our experiment, we have decided to use four discourse link types building on the classification provided by [23] as we hypothesize that the value of semantic similarity might be a useful distinctive factor. The sampled document pairs were classified to the following types: *tangent*, *similarity/equivalence*, *expansion*, *aggregate*. Examples of these link types are depicted in Table I. The description of these link types is as follows:

**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.

**Tangent** links represent according to [23] 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.

| 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 I
EXAMPLE LINK TYPES

**Aggregate** links are those which group together several related documents. According to Allan, 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 thus aggregate links are assigned to document pairs when the first article contains significant parts of the second article.

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.

### D. Results

To answer the questions defined in Section V-B, 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 1 has been divided

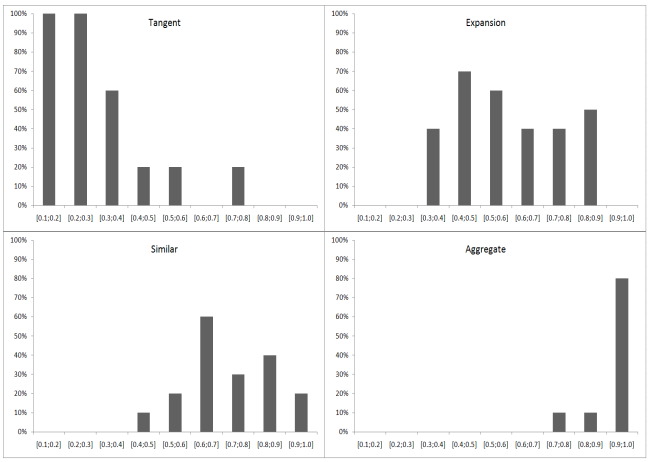


Figure 2. The frequency of different link types with respect to semantic similarity of document pairs

into 9 intervals of even width. As a case study, 10 article pairs from each interval[2] between which a link was created by Wikipedia users were randomly sampled and they were assessed by a human investigator and classified. 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 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.

Overall, 37% of article pairs were classified as *tangent*, 36% as *expansion*, 20% as *similar* and 7% as *aggregate*. The results of the evaluation are presented in Figure 2. The figure shows the frequency of different link types in all the 9 selected intervals.

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 large value of similarity.

Overall, this confirms that the value of semantic similarity is a useful factor characterizing up to certain extent the type of the semantic relationship which provides answer to the second question reported in Section V-B. We have also observed from this experiment and Figure 1 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 which is useful for the detection of certain link types, but has not been used in link typing previously. We expect that robust link typing systems should be developed by combining a number of strategies. 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 *prerequisite* link type, nor it can be used to determine the direction of the link. Other text characteristics perhaps combined with external knowledge should be used for this purpose.

## VI. Conclusion

We have shown that automatic link generation and typing systems are needed in order to provide scalable solutions to document interlinking in large text collections. We argued that cross-document relations cannot be simply produced by the "Social Web" using crowdsourcing methods. However, the automatically identified relations can be confirmed or rejected using social tagging and both approaches can work in synergy.

We have presented an experimental study that shows that the value of semantic similarity is a useful indicator that can help to identify link types. We assume that more similar indicators exist and their combination would improve the accuracy of link typing. In our study, we have used Wikipedia as a source of textual document. This choice alowed us to simplify the problem by considering only a limited set of cross-document relations. In the future, we plan to perform similar experiments on data from scholarly databases that provide more complex and challenging environment for link generation and typing. In addition, we plan to work with lexical units of a lower granularity, such as paragraphs, sentences and noun phrases. This will help us to better understand the characteristics of cross-document relationships with the aim to find distinctive features for

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

various relationship types. This should enable the building of automated and scalable tools for automatic link generation and typing capable of supporting various reasoning and navigation tasks outlined in the beginning of this paper.

## REFERENCES


[1] Knoth, P., Novotny, J., Zdrahal, Z.: Automatic generation of inter-passage links based on semantic similarity. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, Coling 2010 Organizing Committee (2010) 590–598

[2] Henning, V., Reichelt, J.: Mendeley - A Last.fm For Research? In: 2008 IEEE Fourth International Conference on eScience, IEEE (2008) 327–328

[3] Uren, V., Shum, S.B., Li, G., Domingue, J., Motta, E.: Scholarly publishing and argument in hyperspace. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, ACM (2003) 244–250

[4] Shum, S.B.: Cohere: Towards web 2. 0 argumentation (2008)

[5] Burns, G.A.P.C., Cheng, W.C.: Tools for knowledge acquisition within the neuroscholar system and their application to anatomical tract-tracing data. Journal of Biomedical Discovery and Collaboration **1** (2006) 10+

[6] Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The swan biomedical discourse ontology. J. of Biomedical Informatics **41** (2008) 739–751

[7] Wilkinson, R., Smeaton, A.F.: Automatic link generation. ACM Computing Surveys **31** (1999)

[8] Huang, W.C., Geva, S., Trotman, A.: Overview of the inex 2009 link the wiki track. (2009)

[9] Itakura, K.Y., Clarke, C.L.A.: University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks. [30] 132–139

[10] Jenkinson, D., Leung, K.C., Trotman, A.: Wikisearching and wikilinking. [30] 374–388

[11] Lu, W., Liu, D., Fu, Z.: Csir at inex 2008 link-the-wiki track. [30] 389–394

[12] Geva, S.: Gpx: Ad-hoc queries and automated link discovery in the wikipedia. In Fuhr, N., Kamps, J., Lalmas, M., Trotman, A., eds.: INEX. Volume 4862 of Lecture Notes in Computer Science., Springer (2007) 404–416

[13] Dopichaj, P., Skusa, A., Heß, A.: Stealing anchors to link the wiki. [30] 343–353

[14] Granitzer, M., Seifert, C., Zechner, M.: Context based wikipedia linking. [30] 354–365

[15] Allan, J.: Building hypertext using information retrieval. Inf. Process. Manage. **33** (1997) 145–159

[16] Zeng, J., Bloniarz, P.A.: From keywords to links: an automatic approach. Information Technology: Coding and Computing, International Conference on **1** (2004) 283

[17] Zhang, J., Kamps, J.: A content-based link detection approach using the vector space model. [30] 395–400

[18] He, J.: Link detection with wikipedia. [30] 366–373

[19] Trigg, R.: A Network-Based Approach to Text Handling for the Online Scientific Community. PhD thesis (1983)

[20] Mancini, C.: Cinematic Hypertext: Investigating a New Paradigm. (2005)

[21] Shum, S.B., Motta, E., Domingue, J.: Scholonto: an ontology-based digital library server for research documents and discourse. International Journal on Digital Libraries **3** (2000) 237–248

[22] Sanders, T., Spooren, W.P.M., Noordman, L.G.M.: Coherence relations in a cognitive theory of discourse representation. Cognitive Linguistics (1993)

[23] Allan, J.: Automatic hypertext link typing. In: HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext, New York, NY, USA, ACM (1996) 42–52

[24] Marcu, D., Echihabi, A.: An unsupervised approach to recognizing discourse relations. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2002) 368–375

[25] Radev, D.R., Zhang, Z., Otterbacher, J.: Cross-document relationship classification for text summarization. Unpublished paper (2008)

[26] Knoth, P.: Semantic annotation of multilingual learning objects based on a domain ontology (2009)

[27] Shum, S.B., Ferguson, R.: Towards a social learning space for open educational resources. In: OpenED2010: Seventh Annual Open Education Conference. (2010)

[28] Ellis, D., Furner-Hines, J., Willett, P.: 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, New York, NY, USA, Springer-Verlag New York, Inc. (1994) 51–60

[29] Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in wikipedia. In: CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (2007) 453–462

[30] Geva, S., Kamps, J., Trotman, A., eds.: 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. In Geva, S., Kamps, J., Trotman, A., eds.: INEX. Volume 5631 of Lecture Notes in Computer Science., Springer (2009)