EVALUATION OF AUTOMATIC LINKING STRATEGIES FOR WIKIPEDIA PAGES

Michael Granitzer
Graz University of Technology
Inffeldgasse 21a, Austria

Mario Zechner
Know-Center Graz
Inffeldgasse 21a, Austria

Christin Seifert
Know-Center Graz
Inffeldgasse 21a, Austria

Josef Kolbitsch
Graz University of Technology
Steyrergasse 30, Austria

Peter Kemper
Shell International B.V.
Netherland

Ronald In't Velt
Shell International B.V.
Netherland

ABSTRACT

Wikipedia contains an enormous amount of human knowledge. The wide range of covered topics is hierarchically organized in categories and strongly inter-linked. Its structure, its size and the fact that it is generated by humans are the reasons for the attention Wikipedia receives from researchers in different fields. Especially the link structure of Wikipedia is of huge importance not only for humans browsing the collection, but also as a resource for bootstrapping machine intelligence and the semantic web.
Motivated by the fact that manual maintenance and creation of hyperlinks is labor intensive, this paper explores properties for automatic link creation between Wikipedia pages in this paper. Focusing on ad-hoc linking approaches we evaluate linking strategies on the word as well as on the document level using a standard test data set. As it is shown, rather simple approaches yield to reliable results and may be applicable in different application scenarios. Disambiguation strategies based on standard IR techniques help to boost accuracy delivering reasonable results.

KEYWORDS

Retrieval, Link Generation, Evaluation, Wikipedia

1. INTRODUCTION

Wikipedia as a collaboratively created resource of human knowledge contains an enormous amount of human knowledge. This knowledge can not only be exploited by humans, but also used by intelligent algorithms in domains like information retrieval, machine learning or web science. So for example indexing of documents using Wikipedia has been explored in [Medelyan et. al. 08] showing its application for information retrieval.

Similarly, Wikipedia allows to bridge language barriers by being used as a translator in cross-lingual information retrieval [Sorg & Cimiano 2008] or may help to bootstrap the semantic web [Wu and Weld 07].

One important property of Wikipedia articles is their link structure. Users can explore a particular domain easily by following hyperlinks annotated by the author. However, manually annotated links will be incomplete, since due to the sheer size of the collection authors cannot be aware of all related topics. Also, human created links are not only driven by content, but also by beauty and readability. A lot of manually created links are annotated only at the first occurrence of a link target and not on all occurrences in a Wikipedia page. Reducing manual efforts are of lower priority in collaborative environments with large amounts of users, which is shown by the fact that Wikipedia has been created without automatic support. However, for successfully transferring the Wikipedia paradigm of having a central, collaboratively created knowledge source into a corporate setting the support of automatic linking is an essential tool for bootstrapping an enterprise Wiki and for increasing its maintainability. Furthermore, automatic linking of document repositories or other corporate information sources to Wikipedia pages allows for bridging the gap between the heterogeneous information pools of companies.

Application scenarios for automatic creation of hyperlinks are manifold. Besides supporting users in authoring Wikipedia pages automatic linking mechanisms are usable to enhance information access and to semantically enrich information sources by providing links to valid knowledge. Although recent research efforts show increasing interest in automatically linking Wikipedia pages, little is known on the properties for linking. In our contribution we analyze linking properties by investigating different ad-hoc strategies for creating outgoing links (further called "outlinks") on orphan Wikipedia pages. Additionally, our evaluation is based on a standard test data set and thus results are comparable for other research groups.

The remainder of this contribution is organized as follows: Section 2 outlines the state-of-the-art while section 3 introduces our linking strategies. Evaluation and obtained results are discussed in section 4. With the conclusion in section 5 we summarize our findings and point to future work.

2. STATE OF THE ART

Different approaches from different research fields can be found on the particular task of automatically linking documents. Due to the importance as knowledge resources Wikipedia based data sets have recently been in the focus of research. For example the INEX[1] evaluation initiative initiated the so called "Link-the-Wiki" track in 2007 with the goal to explore the automatic construction of links between Wikipedia pages.

One important aspect in this task is the structural level at which links are inserted. On the two extremes, links can be inserted on document-to-document level, which has been done at the INEX challenge, or on word-to-word level, which is one of the tasks in the INEX 2008 challenge.

Document-to-document linking approaches are often based on query by example or content based similarity search techniques. For example [Jenkinson and Trotmann 07] extracted the terms over represented in a document and generated queries of different length. From the BM25 ranked search results the top 250 candidates are considered as link targets. Different to this, in [Geva 07] outlinks are identified by a sliding window approach on the document. The text in the window (1-8 words) is searched for titles containing the text with the GPX XML retrieval engine. The highest ranked pages are chosen while longer page names have been ranked higher. The approach seems to be naive, but results show that it was quite effective.

Similarity measures used have a major impact on the accuracy and runtime behavior of the linking task. In contrast to the usual TFIDF Cosine Similarity calculation, concept mapping techniques map document vectors into a so called concept space. Each dimension of this concept space consists of a particular concept contained in the data set and similarity is estimated within this concept space. Concepts are usually extracted automatically with focus on reducing the problem of homonyms and synonyms as well as reducing the dimensionality of the original vector space.

Latent Semantic Indexing (LSI) also known as Latent Semantic Analysis (LSA) [Deerwester et. al. 91] is one prominent algorithm for concept indexing based on singular value decomposition. Concept clustering, introduced by [Karypis & Han 00], is another technique for creating such concept spaces. Thereby, clustering partitions the data into groups of similar documents optimizing a given loss function. Each cluster may be

[1] INitiative for the Evaluation of XML Retrieval

seen as a concept and vice versa each document or part of a document may be indexed by such clusters through similarity calculation. Thus, similarly to LSI documents are indexed in a concept space but in contrast to LSI the cluster space is usually not dense due to the fact that dimensions (i.e. clusters) are not required to be orthogonal to each other.

In contrast to concept clustering and LSI, Random Projections (RP) are mainly used for reducing the dimensionality of the vector space and do not consider homonymous or synonymous properties of the data set. RPs randomly combine dimensions of the original space and project those dimensions onto a ``new'' dimension. The Johnson Lindenstrauss Lemma [Johnson et. al. 86] provides the rational for using random dimensions, i.e randomly created concepts: it basically states that a sparse vector space can be randomly mapped into a lower dimensional space while preserving the distances. The error of the mapping is bound on the number of data samples in the original space and its dimensionality. Random projections are often used as local sensitive hash functions for detecting near duplicates. Unfortunately random projections are hard to parameterize and yield lower accuracy compared to standard methods [Stein 07], [Salakhutdinov & Hinton 07].

Recently, new approaches on Neuronal Networks, so called deep auto encoders have been explored in [Salakhutdinov & Hinton 07]. Here Boltzmann Machines provide the mapping on the low dimensional space. Evaluations on several test data sets show higher accuracy than for LSI. Surprisingly, the simple TFIDF Cosine similarity achieves similar performance as the Boltzmann Machine.

Creating links on a word based level strongly relates to information extraction tasks like named entity detection. In particular, [Wu and Weld 07] created hyperlinks by matching word sequences with Wikipedia titles and anchor texts. In addition, the use of machine learning techniques like maximum entropy models [Baldridge et. al. 01] may supplement or replace annotations based on manual gazetteers and grammars. However, to the best of the author's knowledge there is no work using machine learning techniques for annotating Wikipedia hyperlinks.

Our contribution focuses on an in-depth analysis of properties for automatically generating links between wiki pages. Similar to [Wu and Weld 07] we focus on sequence matching techniques for annotation on the word level. However, in contrast to them we analyze a set of different matching and disambiguation properties and apply those techniques on the standardized INEX Wikipedia collection. Results are compared to linking strategies on the document level. Also, matching properties have not been analyzed in INEX 07 Link-the-Wiki track and only linking on the document-to-document level has been considered. Therefore, our work intents to provide - in addition to the INEX 07 challenge and the analysis in [Wu and Weld 07] - base line accuracy and comparable results in this domain serving as baseline for more sophisticated linking methods. Also, we investigate different document segmentation levels for disambiguation. Our results underline the findings in [Wu and Weld 07] regarding the usability of such simple ad-hoc approaches in different scenarios.

3. LINKING STRATEGIES

For our experiments we distinguish between linking on word level and on document level. Since correct linking also depends on the context resp. sense of a link, this section also outlines disambiguation strategies used in our experiments.

Word Level Linking: Linking on the word level starts with creating a lookup list, in the following termed as gazetteer, from the Wikipedia collection. Each entry in the lookup list contains a word sequence and the identifier of the Wikipedia page it points to. For creating the lookup list we are distinguishing between taking the title of the Wikipedia page as gazetteer entry and taking title of the page and all anchor texts linking to the page as gazetteer entry. For gazetteer construction only links to other existing Wikipedia pages are considered, while ignoring links to external pages. Matching is done by simply comparing gazetteer entries with the sequence of words in the orphan document. The orphan document is assumed to be a new Wikipedia page or an external document without any link information and thus, matching must rely on the content of the orphan document.

Due to the potentially large size of the gazetteer efficient matching is ensured by the use of finite state machines, similarly as in IE Frameworks like for example GATE (see [Cunningham et. al. 02]). Thereby, transitions between states are words occurring in a gazetteer entry, while states are distinguished into final

states or intermediate states. Final states contain the Wikipedia page and if upon matching such a final state is reached, an annotation pointing to the particular Wikipedia page is added. In this way gazetteer matching allows us to annotate word sequences with hyperlinks for a large number of possible link targets at reasonable speed.

However, the following parameters have been considered during the matching process for evaluation

- Case sensitive vs. Case Insensitive Matching: Since one can assume that a large number of potential matching results will be returned by the matching process, one question is if case sensitive matching yields more reliable results.
- Part-of-Speech Tags: One aspect in matching is the impact of Part-Of-Speech tags on the matching performance. Therefore, orphan Wikipedia pages as well as the gazetteer entries have been preprocessed using the OpenNLP Toolkit [Baldridge et. al. 01] allowing to match sequences only on specific Part-of-Speech tags
- Longest Common Sequence: After gazetteer matching it may happen that annotations overlap or that particular word sequence are annotated by more than one link. One approach to resolve those ambiguities is to consider only the longest matching sequence of words and dropping all others.

Ranking on the Word Level: While gazetteer matching is a binary decision, we are also interested in ranking annotations. The simple ranking based mechanisms outlined in the following should provide a baseline for the ranking obtained from the more sophisticated, context based disambiguation strategies outlined below.

Our first approach, the Inverse Sentence Frequency Ranking (ISFR), borrows the inverse document frequency measure from the well known TFIDF measure and maps it on the sentence level. Similarly for each annotation we are calculating its "inverse sentence frequency" for every document as $w_k = \log\left((n_s + 1)/(n_{s,k} + 1)\right)$ where w_k is the weight for all annotations pointing to page k, n_s is the number of sentences in the current document and $n_{s,k}$ is the number of sentences containing an annotation to page k. Similarly to the link statistic weighting, our rational is to omit high frequency annotations like "The", "Are" etc. since it is very likely that they will occur in a large number of sentences.

The second word level ranking approach, the Inverse Document Frequency Ranking, takes the word distribution over the whole corpus into account. Thereby the number of documents containing the exact text of an annotation divided by the amount of documents in the collection gives the inversed document frequency, where from the logarithm is taken similarly as for ISFR ranking.

Document level linking: While word level linking provides links from a sequence of words of the orphan document to a Wikipedia page, document level linking estimates links to Wikipedia pages for the whole document. One severe difference is that for word level linking a link to a particular Wikipedia page may occur several times while on document level the link only occurs once.

In our experiments nouns are extracted from documents and documents segments using the OpenNLP framework. From the extracted nouns an "OR query" is created and used for searching with the Lucene search engine, which maintains an inverted index over the content of Wikipedia pages. Returned documents are ranked according to their TFIDF score and considered as link target. In case of invoking queries for document segments (i.e. sentences and topics), results of each segment are merged into one result list afterwards. Four merging strategies, namely average score, maximum score, count and average rank are compared.

In contrast to the assumption behind taking the whole wiki page as query source links in Wikipedia pages must not be directly related to the topic of a page. For example the Wikipedia page on "Educational progressivism" contains a lot of different topics ranging from "Sputnik" to "The Boy Scouts of America". It is very unlikely that one query contains all relevant words for a topic and that the results of one query can satisfy all topics. Therefore, we segmented the document on sentence and sub-topic level, whereas sentences are obtained from linguistic analysis using the OpenNLP framework and sub-topics by applying the well known C99 segmentation algorithm [Choif 00]. For each segment a query is constructed and invoked as outlined above.

Disambiguation Strategies: Obviously, word level linking based on gazetteer matching yields a large number of most often wrong annotations. In contrast to the real Wikipedia, the test data set from the INEX collection did not contain any disambiguation pages. Thus, our experiments utilize content based disambiguation over the whole Wikipedia corpus.

In particular, for disambiguation of the automatically generated links word level linking is combined with document level linking. The content of a page is annotated with links on the word level as outlined above. For each segment (i.e. sentence, topic or the whole document) containing a specific link similar Wikipedia pages are searched as described in the document level linking strategies. If the annotated link is now contained in the search result it is accepted as valid link from the source page to the target page with confidence set to the score in the search result lists.

Disambiguation focuses on taking the surroundings of an annotated link into account and to see whether the link occurs in a similar context in the whole collection. By taking different document segments into account we try to exploit the multi-topic nature of Wikipedia pages.

4. EVALUATION

Evaluation has been performed on the INEX 2007 Link-the-Wiki dataset (see [Wei Che et. al. 2007]) with focus on estimating the quality of outlink generation. The data set consists of 659,413 Wikipedia pages, where from 90 topics - one topic is a Wikipedia page - are used as test set. The collection itself is a subset of the Wikipedia XML Corpus described in [Denoyer and Gallinari 2006]. The test set consists of 8,392 outlinks with an average of 94.29 links per document. In total, the test collection links to 5590 unique Wikipedia pages. Gazetteer construction involves all Wikipedia pages excluding the test set resulting in 659,323 gazetteer entries for titles only and around 1.7 million entries by including anchor texts.

Our evaluation focuses on analyzing different aspects of word level annotations as well as on suitable disambiguation strategies. We compare basic word level annotations, disambiguation strategies on the word level as well as document level strategies to derive properties of efficient linking. The TREC evaluation program *trec_val* is used to calculate the different performance measures. In particular estimates for precision (PR) and recall (RE) as well as ranking based measures like Mean Average Precision (MAP) and R-Precision (R-Prec) are provided.

Word Level Linking Results: For word level linking we compared Title (T) vs. Title & Anchor Text (TA), Case Sensitive Matching (CS) vs. Case Insensitive Matching (CI) and Longest Common Sequence Filtering (LCS). Results for word level linking are provided in Table 1. Obviously, matching with anchor texts (TA) leads to a recall close to 1.0 while precision is very low. For a human reader this would yield to mostly all words being hyperlinked to another page. Especially some pages like for example “The” or “Are” are generated very frequently, especially in the case insensitive runs. Considering only case sensitive matches increases precision significantly by a factor of 10, but results are still overwhelming human readers. However, this first analysis shows that focus must be on increasing precision values in order to get usable matches.

Comparing title only matches (T) to title and anchor matches (TA) shows a significant difference in precision and recall. Precision increases by a factor of 25 for the case insensitive baseline runs while recall only drops by a factor of 2. Similarly, considering only the longest common sequence further increases precision and yields in the title only, case sensitive run to the highest micro-averaged precision. We also experimented with taking only words with particular part of speech for matching. However, no significant improvements could be found and therefore results are not presented here. Regarding ranking capabilities Corpus IDF ranking outperforms the others. Noteworthy all ranking mechanisms prefer title only annotations, pointing to the conclusion that noise added by TA matches could not be reduced due to the ranking scheme.

In summary, simple matching strategies without refinement do not provide high enough precision estimates. However, using titles only and filtering out matches based on longest sequence or the simple confidence estimates yields to reasonable results. Also, ranking based on Corpus IDF turns out to provide usable results, while filtering all words but nouns surprisingly stayed below the accuracy of case sensitive matching. Low precision mainly results from stopword gazetteer entries like “The”, “it” etc. and entries which maybe a correct link as for example “Germany”, “scientific theory” etc. but not judged suitable by the author of the Wikipedia page. As outlined in [Wu and Weld 07], it is common that authors only link the first occurrence of a particular word to increase readability. Not detected but correct annotations – responsible for the recall – depend strongly on the chosen parameters. For example for the longest common sequence

parameter it is more likely that annotations like “Austria” will be overwritten by a longer sequence and thus not been detected.

Table 1. Evaluation of word level linking including different ranking schemes

Ranking Mode	**None**		**Inverse Sentence Frequency Ranking**		**Corpus IDF**	
Parameters	**PR**	**RE**	**R-prec**	**map**	**R-prec**	**map**
T, CI (baseline)	0.0741	0.7184	0.1979	0.1469	0.4490	0.3845
T, CI, LCS	0.0987	0.6983	0.2239	0.1623	0.4589	0.3939
T, CS	0.1907	0.5469	0.2918	0.1797	0.3864	0.3156
T, CS, LCS	0.2432	0.5367	0.3250	0.2080	0.4030	0.3247
TA, CI (baseline)	0.0038	0.9682	0.0062	0.0087	0.3598	0.3016
TA, CI, LCS	0.0329	0.6335	0.0778	0.0566	0.3879	0.3044
TA, CS	0.0056	0.9596	0.0083	0.0108	0.3761	0.3148
TA, CS, LCS	0.0341	0.6535	0.0815	0.0614	0.4067	0.3225

Document Level Linking Results: For document level linking we analyzed the influence of segmentation levels, namely document segmentation (Doc), topic segmentation (TO) or sentence annotation (SE) and different merging strategies for those segmentation levels.

Table 2 shows the result for document level linking. Surprisingly, considering different document segmentation levels decreases accuracy; taking the whole document into account yields to the best results. Also, different merging strategies do not increase accuracy for sentences or topics. Regarding the merging strategies clearly average score and average rank performs worse, while count and maximum score achieve a better accuracy for all three measures.

Comparing document level linking to word level linking clearly shows that word level linking achieves higher accuracy estimates for corpus IDF and similar accuracy for inverse sentence frequency ranking. So surprisingly simple matching based approaches combined with local statistics lead to similar results as similarity search approaches. One explanation for that is the locality of Wikipedia links. Often, links are pointing directly to an entity, like “Soviet Union”, without having surrounding words describing the link in depth. Thus, while word level matching clearly identifies “Soviet Union”, the surrounding words do not contain a hint on the link.

Table 2: Evaluation results for document level linking

Document Segmentation Level	***Merging Strategy***	***11-pt avg. Precision***	***R-prec***	***map***
Sentence	Avg. rank	0.0002	0.0005	0.0000
Topic	Avg. rank	0.0087	0.0122	0.0044
Sentence	Avg. Score	0.0164	0.0243	0.0073
Sentence	Max. Score	0.1068	0.1581	0.0793
Topic	Count	0.1130	0.1711	0.0887
Topic	Avg. Score	0.1305	0.1686	0.1030
Topic	Max. Score	0.1711	0.2132	0.1403
Sentence	Count	0.1741	0.2293	0.1413
Document	None	0.2047	0.2524	0.1733

Disambiguation Results: Disambiguation is based on combining word and document level linking. As pointed out by the previous experiments the question in case is whether word level linking accuracy can be further increased by using document level strategies. In our experiments, outlined in Table 3, we focused on the difference on document segmentation levels as well as comparing title with title and anchor matches in order to see whether results of word level linking can be improved. In particular, for every link identified by word level linking a similarity search for this link is invoked, whereas the context is taken with respect to the segmentation level. The score returned by the search engine is used as score for the word level link.

As the results in Table 3 show, again case sensitive matching is central for achieving good results. The noise induced by case insensitive matches cannot be removed at a significant level by using any disambiguation technique. Surprisingly there is no winning document segmentation strategy. Therefore the context of a complete document is sufficient to improve results by disambiguation, which may be due to the topical compactness of a Wikipedia entry compared to for example intranet documents.

Comparing results to document level linking, results clearly improve. However, compared to word level linking, Corpus IDF yields to better results indicating that stop word links like “The” and “Are” are more efficiently removed by global statistics than by disambiguation. For link confidence and Inverse Sentence Frequency ranking and title & anchor (TA) matching benefits most from disambiguation pointing to the conclusion that anchors can be disambiguated efficiently using the proposed strategy.

Table 3. Evaluation of disambiguation strategies

Case sensitivity	*Title only matching*	*Segmentation Level*	*map*	*11-pt. Avg. Precision*	*R-prec*	*Micro Precision*	*Micro Recall*
False	False	Document	0.0892	0.1128	0.1795	0.0871	0.2308
False	False	Topic	0.0902	0.1183	0.1646	0.0176	0.5303
False	False	Sentence	0.1012	0.1255	0.1743	0.0111	0.5916
False	True	Sentence	0.1031	0.1257	0.1930	0.0270	0.4424
False	True	Document	0.2076	0.2367	0.3217	0.2387	0.4769
False	True	Topic	0.2136	0.2456	0.3421	0.2645	0.4525
True	False	Sentence	0.2260	0.2602	0.3358	0.3133	0.3668
True	False	Topic	0.2322	0.2605	0.3378	0.2202	0.5213
True	True	Topic	0.2529	0.2833	0.3681	0.2215	0.4317
True	True	Sentence	0.2529	0.2842	0.3679	0.2244	0.4318
True	True	Document	0.2584	0.2854	0.3539	0.1449	0.5810
True	False	Document	0.2646	0.2907	0.3607	0.1546	0.5783

Comparison to literature: While we did not use the evaluation system of the INEX 07 Link-the-Wiki track, document level results should be comparable. Overall, results on document level linking are staying behind the best results of the INEX 07 Track, which is around a value of 0.484 for Mean Average Precision (see [Geva 2007]). Results on word level would have been ranked fourth in the Link-the-Wiki track. This results are not directly comparable, since our evaluation has been done on a word level not on the document level. It is likely that results improve if our experiments on word level are compared to the document level ground truth.

[Wu and Weld 07] did their experiments on a completely different corpus. However, their results are far higher than our estimates. Arguable factors therefore reside in their approach for only considering link consisting of proper nouns, thereby reducing noise a-priori. In addition, they corrected the ground truth to neglect the human annotation behavior for readability. In their work, all links which are annotated only once by the author, but occurred several times in the document have been corrected manually. Considering their estimate of around 70% human recall, i.e. the number of annotations identified by the author compared to the corrected data set, our precision estimates would be also increased.

5. CONCLUSION

Wikipedia is an enormous resource not only for humans, but also for research fields like information retrieval, machine learning and the semantic web. In this contribution we analyzed automatic creation of links between Wikipedia pages on word and document level as well as analyzed different disambiguation strategies. Our findings can be summarized as follows:

- Word level link approaches based on gazetteer lists yield to reliable results pointing towards direct use in practical settings.
- Simple ranking mechanisms are capable of removing over frequently occurring links like “The” and “Are” and significantly increase accuracy.
- Simple, document centric ranking mechanisms achieve similar results as document level linking, while corpus based ranking mechanisms outperform document level linking.
- Disambiguation strategies significantly enhance disambiguation of anchor texts, but stay in general behind corpus based ranking strategies for word level linking.
- The context around hyperlinks in Wikipedia is usually sparse, indicating the links are used in a more glossary like manner.

Results are also pointing towards two different strategies for increasing accuracy. Since string matching provided good results, machine learning based NLP techniques may further increase accuracy on link detection, especially since Wikipedia provides a huge amount of training data. Also, using concept based disambiguation techniques rather than term based retrieval methods may further increase disambiguation accuracy.

ACKNOWLEDGEMENT

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.
Furthermore, results presented in this paper have been partially funded by Shell International B.V., The Hague, The Netherlands.

REFERENCES

Baldridge, J.; Morton, T. & Bierner, G., 2001, OpenNLP: The Maximum Entropy Framework, Web Site http://maxent.sourceforge.net/about.html, last visited June 2008.

Choif, F. Y. Y., 2000, Advances in domain-independent linear text segmentation, in *Proceedings of the North American Chapter of the ACL*. Seattle, Washington. ACM International Conference Proceeding Series

Cunningham, H.; Maynard, D.; Bontcheva, K. & Tablan, V. , 2002, GATE: A framework and graphical development environment for robust NLP tools and applications, *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Denoyer, L. & Gallinari, P., 2006, The Wikipedia XML Corpus, *SIGIR Forum*.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W. & Harshman, R. A. , 1990, Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science* 41 (6), 391-407 .

Geva, S. (2007), GPX@INEX2007: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia, *Lalmas; Andrew Trotman; Norbert, Fuhr; Mounia, ed.,Pre-Proceedings of INEX 2007*

Jenkinson, D. & Trotman, A., 2007, Wikipedia Ad hoc Passage Retrieval and Wikipedia Document Linking, *Lalmas; Andrew Trotman; Norbert, Fuhr; Mounia, ed.,Pre-Proceedings of INEX 2007*

Johnson, W. B.; Lindenstrauss, J. & Schechtman, G. (1986), Extensions of lipschitz maps into Banach spaces, Israel *Journal of Mathematics* 54(2), 129-138.

Karypis, G. & Han, E., 2000, 'Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization', *Technical report*, University of Minnesota.

Medelyan, O.; Witten, I. H. & Milne, D., 2008, Topic Indexing with Wikipedia., *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*.

Salakhutdinov, R. & Hinton, G., 2007, Semantic Hashing, *Proceedings of the SIGIR Workshop on Information Retrieval and Applications of Graphical Models*, Amsterdam.

Sorg, P. & Cimiano, P., 2008, Enriching the crosslingual link structure of Wikipedia - A classification-based approach, in *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*. June 2008.

Stein, B., 2007, Principles of Hash-based Text Retrieval, in Noriko Kando Wessel Kraaij Charles Clarke, Norbert Fuhr & Arjen de Vries, ed.,*30th Annual International ACM SIGIR Conference*, ACM, , pp. 527-534.

Wei Che, Huang; Yue, X. S. G., 2007, Overview of INEX 2007 Link the Wiki Track, *Lalmas; Andrew Trotman; Norbert, Fuhr; Mounia, ed.,Pre-Proceedings of INEX 2007*, pp. 350—364.

Wu, F. & Weld, D. S., 2007, Autonomously semantifying wikipedia, *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, New York, NY, USA, pp. 41--50.