# Experiments and Evaluation of Link Discovery in the Wikipedia

Wei Che (Darren) Huang
Faculty of IT
Queensland University of Technology
Brisbane, Australia
*w2.huang@student.qut.edu.au*

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
*andrew@cs.otago.ac.nz*

Shlomo Geva
Faculty of IT
Queensland University of Technology
Brisbane, Australia
*s.geva@qut.edu.au*

## ABSTRACT

Collaborative knowledge management systems such as the Wikipedia are becoming ever more popular – and these systems typically contain hypertext links between documents. The Wikipedia offers both manual and automated link creation. In fact several different systems providing links for Wikipedia documents now exit. Problematically the quality of automatically generated links has never been quantified. An evaluation method for Wikipedia link discovery approaches is essential.

We introduce the Link-the-Wiki task launched at INEX in 2007. 90 documents were orphaned from the collection and participants were required to build systems that identified the missing links. The different automated link discovery techniques used by participants are outlined. Details of two successful techniques are given, one using the titles of pre-existing documents to identify anchors and destinations, the other using pre-existing links between documents to identify possible links in new documents. In this paper, we mainly focus on the analysis and assessment of Wikipedia link discovery and discuss possible future evaluation techniques.

We examine one system in further detail and conduct a scalability experiment in which 1% of all Wikipedia documents were used and the performance studied in detail – link discovery in this system is shown to be scalable.

Finally, potential research directions for link discovery, assessment and evaluation are discussed.

**Categories and Subject Descriptors:** H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

**General Terms:** Measurement, Experimentation

**Keywords:** Wikipedia, Link-the-Wiki, INEX, Assessment, Evaluation, Information Retrieval, XML IR

## 1. INTRODUCTION

### 1.1 Motivation

The goal of collaborative hypertext knowledge management (as seen, for example, in the Wikipedia) is to interlink all related knowledge. This helps users realize their particular information need, regardless of their level of understanding, by allowing them to click between text of entries expressing different and related concepts at different and related depths of coverage. Without hypertext links a user must search and browse (or otherwise navigate) into requisite content in order to expand their understanding. It is utterly inconvenient for the user to repeatedly search the collection simply to find content related to their core need. Worse, sometimes the content is not easily reachable using navigational facilities provided by the knowledge management system.

Links between pages are essential for navigation, but most systems require authors to manually identify each link. Authors must identify both the anchors and the target page in order to place a link. This creates a heavy and often unnecessary burden on content providers [1] who should focus on the content and not on the relationship between their content and content already in the collection. As the size of the collection increases the task of manually identifying links can become unmanageable. The maintenance cost of keeping all links up to date is huge – and the Wikipedia has seen faster than linear growth for many years. Authors are typically unaware of all pre-existing content to which they might links, and even if they are they are unlikely to be aware of content created concurrently with their page. Page maintenance, in particular, linking to content added after a page is created is a burden to content providers who often do not maintain their content (hence the collaborative nature of these information resources). Worse, Ellis et al. [2] have shown significant differences in the links assigned by different people.

Several systems (such as the Wikipedia) support simple text-search facilities to help content providers identify anchors and links. External search engines such as Google, Qwika, Lycos and Yahoo can also be used to search well established knowledge management sites [3].

There are further problems! Linking is still typically performed between documents even though some documents are long and a better destination might be an anchor within a document. Link discovery methods have not yet been integrated into even the most successful systems (such as the Wikipedia). Links outside the closed system (i.e. to the web) are also manually added. There are many inaccurate and unnecessary links added to documents. And link spam is beginning to surface.

To eliminate the human effort required to build a highly accurate hyperlink-link network, to reduce the chance of erroneous links, and to keep links up-to-date, automatic link discovery mechanisms are needed.

Herein we concentrate on the Wikipedia because of its success and because of the availability of the INEX Wikipedia document collection. In particular we discuss the Link-the-Wiki track held at INEX 2007 in which automated link discovery systems were

*SIGIR 2008 workshop on Focused Retrieval*
*July 24, 2008, Singapore*

solicited from participants and judged against human created hypertext links for 90 documents.

The techniques used by each participant are discussed and contrasted, and then the results of the top two performing groups are analyzed in detail. We find that document-to-document link discovery systems are very good at exhibiting high precision levels at most points of recall, systems are scalable and that several different techniques might be used. This result motivates us to examine (and outline future work in) anchor to Best-Entry-Point (BEP) identification. We discuss assessment and evaluation of this new focused retrieval task is in detail.

## 1.2 Related Work

As suggested by Wilkinson & Smeaton [1], navigation between linked documents is a great deal more than simply navigating multiple results of a single search query, linking between digital resources is becoming an ever more important way to find information. Through hypertext navigation, users can easily understand context and realize the relationships of related information. However, since digital resources are distributed it has become difficult for users to maintain the quality and the consistency of links. Automatic techniques to detect the semantic structure (e.g. hierarchy) of the document collection and the relatedness and relationships of digital objects have been studied and developed [4]. Early works, in the 1990s, determined whether and when to insert links between documents by computing document similarity. Approaches such as term repetition, lexical chains, keyword weighting and so on were used to calculate the similarity between documents [5, 6, 7]. These approaches were based on a document-to-document linking scenario, rather than identifying which parts of which documents were interrelated.

Several conferences and workshops (in particular at SIGIR and LinkKDD) focused on link analysis and discovery. Most recently the Link-the-Wiki track at INEX required participants to build systems that discover potential anchors (representing the content of topics) and relevant destinations (Best Entry Points within a document) for each anchor [8, 9]. The details of this track are briefly described in Section 2.

The link-network within Wikipedia is only valuable if it is maintained and all links are up-to-date, this is especially a problem in the case of a newly created article that should be linked to by pre-existing pages. Links in each document can be within the Wikipedia or other web resources outside the Wikipedia [10] so the document collection can never be closed. Although there are many methods in modern IR that can be applied to facilitate search, few experiments have been done in collaborative semantic linking [11].

Adafre & de Rijke [12] identify most links in the Wikipedia as conceptual. The Wikipedia link-network offers hierarchical information and links aim to expand the concepts in their anchors. The anchors imply the concept while the links are complementary to the concept. Since there is no strict standard of editing there are problems with *over linking* and *missing links*. Adafre & de Rijke proposed a method of discovering *missing links* in Wikipedia pages by clustering topically related pages using LTRank and identified link candidates by matching anchor texts. Page ranks using the LTRank method are based on the co-citation and page title information. Experimental results showed a reasonable outcome.

Jenkins (2007) developed a link suggestion tool, *Can We Link It*. This tool identifies anchors within a document that have not been linked and that might be linked to other pages [13]**.** Using this tool, the user can accept, reject, or *"don't know"* to leave a link as undecided. This tool also lets the user add links back to Wikipedia document.

A collaborative knowledge management system, called *PlanetMath*, based on the Noosphere system has been developed for mathematics [14]. It is encyclopedic, (like the Wikipedia), but mainly used for the sharing of mathematical knowledge. Since the content is considered to be a semantic network, entries should be cross-referenced (linked). An automatic linking system provided by Noosphere employs the concept of conceptual dependency to identify each entry for linking. A classification hierarchy used in online encyclopedias is used to improve the precision of automatic linking. In practice, the system looks for common anchors that are defined in multiple entries and creates links between them, once the page metadata is identified as related. Based on the Noosphere system, NNexus (Noosphere Networked Entry eXtension and Unification System) was developed to automate the process of the automatic linking [15]. This was the first automatic linking system to eliminate the linking efforts required by page authors. Declarative linking priorities and clauses are specified to enhance linking precision. An approach, called *invalidation index*, was developed to invalidate entries belonging to those concepts where there are new entries. Reputation based collaborative filtering techniques could be used to provide personalized links.

Research on the Wikipedia has been undertaken in recent years. In order to find cultural biases, network analysis algorithms such as HITS and PageRank have been used [16]. Based on Markov Chains [17], a set of experiments for finding related pages within the Wikipedia collection was undertaken using two Green-based methods [18], *Green* and *SymGreen*, and three classical approaches, *PageRankOfLinks*, *Cosine with tf-idf* and *Co-citations*. The results show the Green method has better performance at finding similar nodes than only relying on the graph structure. Although page titles and category structure can be used to classify documents, properties such as the internal text of the articles, the hierarchical category, and the linking structure should be used [19]. *Wikirelate* proposed by Strube & Ponzetto [20] uses *Path*, *Information content* and *Text overlap* measures to compute the semantic relatedness of words. These measures mainly rely on either the texts of the articles or the category hierarchy. Gabrilövich & Markövitch [21] introduce a new approach called Explicit Semantic Analysis (ESA), which computes relatedness by comparing two weighted vectors of Wikipedia concepts that represent words appearing within the content. Common to this research is the use of the *existing* linking structure and content (category, etc.); we are interested in developing approaches to generate *new* links.

Various link-based techniques based on the correlation between the link density and content have been developed for a diverse set of research problems including link discovery and relevance ranking [12]. Moreover, communities can be identified by analyzing the link graph [22]. Beside co-citation used by Kumar et al. [23] to measure similarity, bibliographic coupling and

SimRank based on citation patterns, and the similarity of structural context (respectively), have also been used to identify the similarity of web objects [24]. The companion algorithm derived from HITS has also been proposed for finding related pages (by exploiting links and their order on a page) [25, 26].

The assessment of results has been a challenge in IR experiments for many years because there is no standard procedure, relevance is hard to define and cross-assessor agreement levels are often low (so individual judgments come under dispute). Worse, it is difficult to compare IR methods which are able to retrieve highly relevant documents with those that retrieve less relevant documents because assessments are usually binary. The use of Precision-Recall curves is typical in IR; however, Schamber [27] argues that traditional P-R based comparison using binary relevance cannot adequately capture the variability and complexity of relevance. Relevance is a multilevel circumstance where, for a user, the degree of relevance may vary from document to document.

Several studies have examined components that influence judgments and the criteria of relevance (including graded relevance) in information seeking and retrieval [28]. Kekalainen and Jarvelin [29] argue that evaluation methods should be flexible enough to handle different degrees of judgment scales. They proposed generalized precision and recall that can incorporate a continuous relevance scale into the traditional precision and recall measures. Their experiments demonstrate that the evaluation approach can distinguish between retrieval methods fetching highly relevant documents from those retrieving partially relevant documents.

## 2. INEX LINK-THE-WIKI TRACK

The Wikipedia is composed of millions of interlinked articles in numerous languages and offers many attractive features for retrieval tasks [30]. The current INEX Wikipedia collection contains a snapshot of the Wikipedia English collection from 2006 and contains 660,000 documents and is about 4GB in size. In INEX 2007 the linking task used 90 documents (topics), nominated from the existing collection by participants [31]. Topic nomination was preferred over random selection because some documents contain very few links (because, for example, they are very short). The topic-documents were removed from the collection (as were links to and from the documents) and treated as if new. The task was to identify a set of incoming and outgoing links to and from these orphaned documents together with the corresponding anchor text within the orphaned documents.

### 2.1 Assessment and Evaluation

There are two challenges in the LTW track at INEX. The first is to identify a set of text anchors that may semantically be linked to other pre-existing documents – these are candidate outgoing links. The second is the identification of candidate incoming links from other Wikipedia pages into the new document. Several natural language use issues such as synonymy and multiple meanings may cause anchor text inaccuracies and deficiencies. For example, the term *IR* can mean *Information Retrieval* or *Information Registry*, depending on context. Should *Modern Information Retrieval* or just *Information Retrieval* be highlighted as a term when both articles exist in the collection? As an aside, of course both could be linked, but unfortunately the Wikipedia interface (the web) does not currently (easily) support multiple links per anchor.

It is important to rank the discovered links for a user's selection, but it is not immediately clear how this should be done. A typical scenario might involve a user who wishes to inspect and then accept or reject recommended links. The user is unlikely to go through hundreds of potential anchors. Therefore, the most likely anchors should be presented first. Furthermore, even with automated linking the system must balance extensive linking against link quality. Ranking is necessary in order to determine a cut off point in link recommendation.

It is essential to define a standard methodology and procedures to assess link quality and to quantitatively evaluate and compare different approaches.

INEX 2007 used a variation of the Cranfield methodology. We have already discussed topic selection. From the topics we automatically generated the assessments (the ground-truth). Because the topic-documents were extracted from the existing Wikipedia collection, links both into the collection and from the collection already exist. These were used as the ground-truth, and were then eradicated from both the topic and the collection before the topics were distributed. This ground truth was not ideal (as we shall discuss later) but nonetheless reasonable, as it was what that was in the Wikipedia at the time.

Constructing the assessments in this way resulted in no manual assessment effort, and facilitated the evaluation of systems with a very large number of topics. Participating search engines were explicitly forbidden from using the existing links to and from the topics (the whole collection was used in the INEX ad hoc track complete with links) although links within the collection that were unrelated to the topics could be used. Participant's search engines returned ranked lists of possible incoming and outgoing links for each topic. Evaluation was carried out using MAP, R-Prec and P@R. Incoming and outgoing links were evaluated separately.

This kind of automatic generation of link-assessments is applicable only to document-to-document link discovery because these are the only kinds of links that exist within the collection. Because of this INEX 2007 limited link discovery to document-to-document linking.

The goal of the task is to perform focused retrieval. That is, to link anchors in one document to focused units (e.g. *sections*, *images*, *elements*, or *passages*) in another. An anchor link click should ideally lead a user not only to a relevant document, but also to the best entry point within that document with respect to the anchor context. This requires far more elaborate assessment and evaluation and is discussed later in this article.

### 2.2 The Quality of Wikipedia Links

Although we treat the Wikipedia links as the ground-truth, they are obviously not perfect. Some links in the Wikipedia are already automatically generated and the validity is questionable. *Year* links, for example, are very often unrelated to the content of the document, but are easy to discover. Problematically they may also lead to optimistic evaluation results when identified by link-discovery systems using automatic assessment generation techniques such as we describe. Many potentially good links that have not been identified by Wikipedia users are amenable to automatic discovery – but will not be scored using automatic assessment generation. Such useful returned links which are missing from the ground truth could result in poor evaluation scores for highly effective link discovery systems, leading to pessimistic evaluation results. So although it is not possible to quantify the absolute performance using automated assessment, the procedure we used provides a trade-off between assessment effort (essentially none) and absolute accuracy of measurement.

It is a reasonable to conjecture that *comparative evaluation* of methods and systems is still informative. Through *comparative* analysis of automated linking systems, it should remain possible to improve link discovery methods.

## 3. WIKIPEDIA LINKS

Links in the Wikipedia can be classified into several types. Crudely, they can be divided into linking within Wikipedia and outside web links. Less crudely:

- Linking to an article which has the exact same name as an anchor.
- Linking to an article which has a different name from the anchor, we identify the following kinds:
    - **Synonyms**. Linking to a page whose name has the same meaning as the anchor but different spelling. For example, the word "gods" in the following sentence, *The elves were originally imagined as a race of minor nature and fertility gods*, is linked to the page named *Deity*.
    - **Tense**. The past tense of a word may be linked to a page name as its present tense or its noun form. For example, the "pluralized" in the sentence, *Elf can be pluralised as both elves and elfs*, is linked to the page name *plural*.
    - **Presenter**. A name of an entity may link to its related presenter such as the singer of a song or the director of a film.
    - **Language**. Some old language characters (e.g. Latin and Old Norse) may be linked to related English words. For example, the word "*Ljósálfar*" in the sentence, *he also based them on the god-like and human-sized Ljósálfar of Norse mythology*, is linked to the page *Light elf* which is in turn redirected by Wikipedia to a page titled *Light elves*.
    - **Definition**. Some anchors are linked to the related page names that may express the meaning of the anchor. For example, the word "good" in the following sentence, *They are great smiths and fierce warriors on the side of good*, is linked to the page "*Goodness and value theory*" with the title "*value theory*".
    - **Disambiguation**. Some links are redirected to a page that lists possible linking candidates. For example, the anchor "Moving Pictures" is linked to the page *Moving Pictures* that lists a serious of related pages (e.g. *Moving Pictures (album), Moving Pictures (novel), Moving Pictures (song)* and *Moving Pictures (band)*).
- An anchor may be linked to a page that integrates several similar pages. Anchors that link to these "similar" pages are later redirected to the new integrated page.
- Anchors may be in the *references* section: these are anchors that link to destinations either inside or outside Wikipedia.
- Anchors in the *See Also* section: this is a list of related topics that link to Wikipedia pages.
- *External Link* Anchors: there is also a list of related topics that link to pages outside Wikipedia (that is, to the web).

Problematically, if most page names exactly match an anchor text, we can produce a simple method that systematically matches potential anchor strings with page names to identify most links – and achieve a recall of near 1.

We examined the 90 LTW topics from INEX 2007 and found that in 81 of the 90 topics at least 50% of the links match an existing page name (see Table 1). This could be because the links were generated through careful construction by a user, or automatically by matching page names, either way such links are relatively easy to find. Although this implies that we can expect high recall from simple page-name matching strategies, it does not necessarily mean that we can expect high precision – many matching links are not relevant (for example, polyvalent terms). As the Wikipedia is a huge repository of definitions it is relatively easy to find matching page names which are not relevant.

| Ratio of Match | Number of Topics |
|---|---|
| 90% ~ 100% | 1 |
| 80% ~ 90% | 8 |
| 70% ~ 80% | 26 |
| 60% ~ 70% | 35 |
| 50% ~ 60% | 16 |
| 40% ~ 50% | 2 |
| 30% ~ 40% | 2 |

**Table 1: Ratio of matching names between anchors and links**

## 4. APPROACHES TO LINK-THE-WIKI

In this section we briefly describe the approaches that were taken by the Link-the-Wiki participants.

The University of Amsterdam system assumed that Wikipedia pages link to each other when articles are similar or related in content. For each of the 90 topics, the system queried the index of the entire collection, (excluding the topics). This was done by using the full topic as the query, but excluding stop words, and with important terms derived from a language model. The top 100 files (anchors) were selected for each topic. They experimented with line matching from the orphans to the anchor files. For the outgoing links, the system matched each line of a topic with the lines of the anchors until a matching line was found. For the incoming links, the system iterated over all lines of each anchor for each line of the topic. The generated runs were based on the names of the pages, exact lines, and longest common substrings (LCSS) expanded with WordNet synonyms. The results show that the run based on restricting the line matching to the names of pages performed best.

The University of Otago system identified terms within the document that were over represented by comparing term frequency in the document with the expected term frequency (computed as the collection frequency divided by document frequency). From the top few over-represented terms they generated queries of different lengths. A BM25 ranking search engine was used to identify potentially relevant documents. Links from the source document to the potentially relevant documents (and back) were constructed. They showed that using 4 terms per query was more effective than fewer or more. The Otago system was effective at early recall but not overall.

The University of Waterloo system found the first 250 documents (in document collection order) that contain the topic titles and then generated article-to-article Incoming links. For outgoing links, they performed link analysis. The system computed the probabilities that each candidate anchor would be linked to a destination file. The probability that a candidate anchor would be linked was computed (essentially) as the ratio of the number of times that the anchor text was actually linked in the collection, to the number of times that the anchor text appeared in the collection.

The Queensland University of Technology (QUT) system identified incoming links using a ranked search for documents that were about the new document title. Outgoing links were identified by running a window over the new document text and looking for matching document titles in the collection. The window size varied from 12 words down to 1 word, and included stop words. Longer page names were ranked higher than shorter page names, motivated by the observation that the system was less likely to hit on a longer page name by accident.

The best performing approaches were those that used either existing anchors to predict suitable anchors (Waterloo), or matching document titles to predict suitable anchors. The performance of these 2 approaches[1] is depicted in Figure 1. Both approaches produce a very good result with high precision over a wide range of recall levels. This is precisely the kind of performance needed to satisfy a user.

# 5. EVALUATION RESULTS

In this section we concentrate on the two most successful approaches at INEX 2007 [31, 32], those of Waterloo and QUT.

## 5.1 Anchor vs. Page Title Link Discovery

There are considerable differences between the two approaches. The Waterloo approach relies on the availability of an extensive pre-existing web of anchor to document links in the collection. This pre-requisite may not always be satisfied, particularly when a new cluster of documents in a new domain is added to the collection in bulk, or when a new Wikipedia-like resource is created. However, the approach can discover links that are not solely based on a match between anchor text and a document title. If an anchor is frequently linked to a document with a different title, it will become a highly probable link. For instance, the Waterloo system was able to link *Educational Philosophy* to a document titled *The Philosophy of Education.* By contrast, the QUT approach only discovered matching document titles. Although the performance of QUT is somewhat lower, the approach is applicable to any collection, regardless of the pre-existing link structure. It could immediately be applied to any document collection, new or pre-existing.

Figure 1 presents the precision-recall curves for the two systems. "Anchors 90" is the Waterloo system and "Page Titles 90" is the QUT approach. Both are shown for the 90 INEX topics. The anchor-based approach is better at almost all recall points.

## 5.2 Scalability of Link Discovery

To test the scalability of automated link discovery we additionally ran an extensive experiment on the collection. We randomly extracted 1% of the 660,000 documents and re-ran the experiment. So-far only QUT have provided results.

The QUT experiment was run on a PC with 2GB memory and 1.6GHz clock speed. It took 6 minutes to complete the process, processing in excess of 1,100 documents per minute. Figure 1 also presents the recall-precision curve for that run. It can be seen that performance over a very large number of topics selected at random is similar to the performance achieved over the INEX set, suggesting that 90 topics is sufficient to measure the performance of such systems. This result suggests that the manual choice of topics for INEX 2007 was not biased – which further suggests that topics can be randomly chosen in future years (thus further reducing the cost of assessing such systems for a document-to-document linking scenario).

Importantly, it is feasible to manually assess 90 topics whereas it is not be feasible to assess 6,600 using the resources available to INEX. Manual assessment would allow us to study more deeply the nature of link discovery – to identify those links returned by automatic systems that have not been identified by Wikipedia authors. It would also allow us to identify links that are already in the Wikipedia but which are not useful (e.g. *year* links are common, yet often of little use).

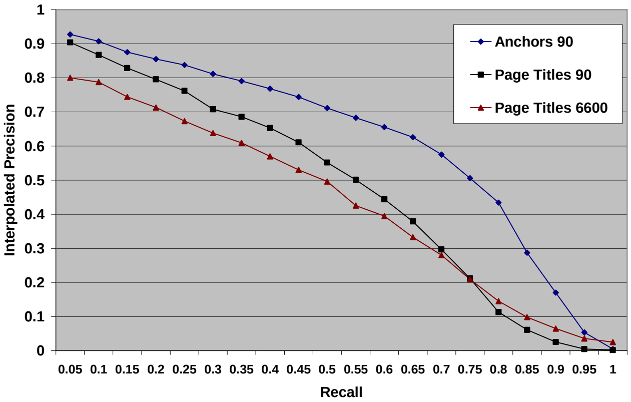

**Figure 1. Scalability test: Differences in performance topic sets is likely to be human bias in topic choice**

## 5.3 Page Name Based Link Discovery

It is straightforward to obtain candidate anchors by systematic comparison of substrings (of various lengths) against exiting page titles in the collection. Numerous matches arise, not all useful, so

[1] The graphs shown in this paper for the participating systems were generated after INEX 2007 and after the participants had fixed bugs and implemented corrections. The results we present will, therefore, not match those reported at INEX.

a pruning strategy is needed. QUT adopted the following (effective) strategy:

- Identify all candidate phrase-anchors of length 12 words down to 2, in that order.
- Append candidate year anchors
- Append all single term anchors

No ordering was performed other than the above. Phrases were ordered by length, followed by years, followed by single terms. Within these groups the ordering was in the sequence in which the anchors were encountered.

QUT found that it is possible to improve their result by re-ordering the combined single-term and year anchors by the probability of the word being an anchor. This probability is estimated as the ratio of the number of times that the named page had been linked to, to the number of times that the page name appears in the collection (the collection frequency). Alternatively they used the number of documents in which an anchor text appears (document frequency). The performance degraded when phrase anchors were used in re-ordering – it appears that the phrase match heuristic is more useful than the estimated phrase anchor probability. Only short single term candidate anchors were ordered by the probability of being linked.

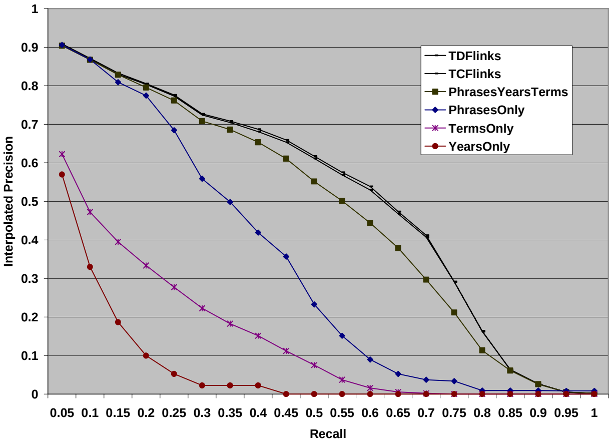

**Figure 2: Linking strategy comparison**

In order to assess the contribution of each component (phrase, year, and term), QUT created separate submissions for each component. Figure 2 presents the recall-precision curves. Most surprisingly the contribution of the year links is small; they are ubiquitous throughout Wikipedia and were expected to contribute considerably to performance. Single terms contribute more than years over all because there are many more terms that could be linked. However, it is difficult to avoid irrelevant links using only single terms. The phrase links achieve higher precision and recall than terms and years. This is because phrases (long phrases in particular) that match a page name are highly unlikely to occur in a document without also being related to the page of the same name. Both years and single terms frequently match a page name, but not in the correct context. The combination of phrases, years, and terms is very effective as can be seen from the combined curve. The ranking of single terms by probability of being an anchor provides further improvements. The top 2 curves in Figure 2 correspond to these variations. The improvement is only marginally greater when using the document frequency in place of collection frequency.

# 6. FOCUSED LINK DISCOVERY

The INEX Link-the-Wiki track in 2007 called for document-to-document linking. The goal of the task is to find links that point not only to a relevant document, but also to a location within that document from which a user should start reading in order to satisfy their information need. This location is called a Best Entry Point, or BEP [33]**.**

Furthermore, it is reasonable to expect to see anchors that could point to multiple locations. For instance, there may be numerous pages about *Education Theory*, not just a single overview page by that name. It may be necessary to impose some limit on the number of links per document, (and the number of links per anchor) to avoid linking every word of every document to another document. Just because the Wikipedia (or rather, a standard web browser) currently does not support the presentation and handling of multiple links per anchor it does not mean that we cannot or should not explore this scenario.

In future INEX evaluations the task will be defined as anchor to BEP link discovery, and allow multiple links per anchor (actually, the latter is essential for manual evaluation purposed where two systems might link the same anchor to different document, both of which are relevant). Traditional performance measures such as MAP (Mean Average Precision) will be adapted to address the performance differences of link-discovery methods in this new scenario.

Automated generation of assessments (the INEX 2007 model) produced an incomplete and biased ground truth, biased on what users did, not what they might (or should) have done. This bias is not dissimilar from that seem with relevance feedback experiments in which a user can only improve on results they have seen, and is not able to identify better and more relevant documents they have not seen. This problem was already explored in Section 2.2. Furthermore, the Wikipedia links do not have anchor-to-BEP functionality, nor do they have multiple links per anchor. Therefore it will not be possible to automatically generate assessments for evaluating anchor-to-BEP runs. For this it is necessary to employ a manual assessment procedure as well as a revised evaluation strategy. We identify this as a necessary future direction for research, and currently study it. An assessment tool for facilitating effective inspection and binary relevance judgment of individual anchor-to-BEP links is outside the scope of this paper, however we also currently study this. In what follows we assume that such a tool exists and that such relevance judgments of links will be available.

## 6.1 Proposed Evaluation Procedure

In automated link discovery there are two simultaneous ranking requirements: first a candidate list of anchors, second a candidate list of target documents for those anchors. In order to derive a single performance score over all proposed anchors and targets, the performance of each must be combined.

A suitable form for a document score may be:

$$A = MAP(links) \qquad (1)$$

Where $A$ is the single anchor score, defined as the mean average precision. For evaluation purposes runs will be of finite length so MAP will be computed up to that point of recall.

We must also allow for anchors to be matched with some flexibility. An anchor may be defined in several slightly different ways. For instance, *The Theory of Relativity*, *Theory of Relativity*, and *Relativity* may well be conceptually identical anchors. Furthermore, if the anchor text occurs several times in a document we would expect only one instance to be anchored (as, for example, is seen in the Wikipedia) and so the location of the anchor may vary without being logically incorrect (we leave for yet-further work the question of which occurrence of an anchor is best to choose). In deriving a relevance score for an anchor a match has to be defined as conceptual, requiring only some minimal term overlap with an anchor in the assessments. The same kind of problem is seen in Question Answering where templates have been used to match correct answers, rather than document locations.

Similarly, a BEP cannot be defined with absolute accuracy. Some reasonable proximity to a designated BEP in the assessments should be allowed. So a BEP might be considered relevant if, when viewed on a screen, it is no more than some distance (*N* words) away from a point chosen by an assessor (INEX uses a similar scheme for scoring BEPs).

So in summary, an anchor-to-BEP link can be assessed as relevant on the basis of approximately matching both the anchor and the BEP of a relevant link in the assessments.

Having computed individual anchor-to-BEP link score the document score can be derived:

$$D = \sum_{anchors} f(P)_i A_i \qquad (2)$$

Where $A_i$ is the score assigned to a particular anchor, and $f(P_i)$ is a monotonically decreasing function of the position of the BEP in the target document. The score can then be averaged over all topics in a run to provide the final run score.

## 7. CONCLUSIONS AND FUTURE WORK

As far as we are aware, the Link-the-Wiki task at INEX is the first to offer extensive reusable independent evaluation resources for link discovery. We have described this new evaluation task and then compared and contrasted the two most successful approaches submitted to Link-the-Wiki at INEX 2007. We further provided results of extensive linking experimentation with a very large set of documents (1% of the collection) and found that linking is feasible, effective, and scalable.

A fully automated procedure for document-to-document link analysis that costs virtually nothing to administer is described. The procedure was used at INEX 2007 and allowed us to create a fast evaluation procedure with a turnaround time of days and not months because no manual assessment was required. The procedure allows for a very large number of documents to be used in experiments, and we demonstrate this by using 6,600 documents for assessment. For link-discovery we have overcome the assessment bottleneck which is encountered in most other tasks in collaborative evaluation forums such as INEX and TREC. We further proposed to extend the task to anchor-to-BEP link discovery, and to multiple links per anchor. We describe the requirements for evaluating such a task and propose an evaluation procedure that is derived from standard well established IR methodology of measuring MAP.

There is still much to explore in link discovery in Wikipedia. For document-to-document link discovery there was no demonstrated successful use of document similarity metrics to determine the appropriateness of a link. Sub-document similarity measures (as seen in ad hoc focused-retrieval experiments) are expected to be successful for BEP identification. That is, the similarity between the immediate anchor's context and the immediate BEP context. It is not necessary for whole documents to be highly related for valuable links to exist. For instance, a document on Information Retrieval which briefly refers to Latent Semantic Analysis may well link to a document which discusses Dimensionality Reduction. Although the two *documents* may not seem related, a *section* on latent semantic analysis in one document may link to a *section* on singular value decomposition in the other. There is ample scope for natural language processing technology to explore ways by which context similarity can be used to improve the accuracy of link analysis at a granularity well below whole document.

We believe we have solved the evaluation problem for document-to-document linking and currently explore the evaluation of anchor-to-BEP linking in the context of focused-retrieval.

## 8. REFERENCES

[1] Wilkinson, R. and Smeaton, A. F. (1999) Automatic Link Generation, *ACM Computing Surveys*, 31(4), December 1999.

[2] Ellis, D., Furner-Hines, J. and Willett, P. (1994) On the Measurement of Inter-Linker Consistency and Retrieval Effectiveness in Hypertext Database, *In Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994, 51-60.

[3] Wikipedia: Searching, 2007. http://en.wikipedia.org/wiki/ Wikipedia: Searching.

[4] Green, S. J. (1999) Building Hypertext Links By Computing Semantic Similarity, *IEEE Transactions on Knowledge and Data Engineering*, September/October 1999, 11(5), 713-730.

[5] Allan, J. (1997) Building Hypertext using Information Retrieval, *Information Processing and Management*, 33(2) 145-159.

[6] Green, S. J. (1998) Automated Link Generation: Can We Do Better than Term Repetition?, *In Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 75-84.

[7] Zeng, J. and Bloniarz, O. A. (2004) From Keywords to Links: an Automatic Approach, *In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*, 5-7 April 2004, 283-286.

[8] Denoyer, L. and Gallinari, P. (2006) The Wikipedia XML Corpus, *ACM SIGIR Forum*, 40(1), June 2006, 64-69.

[9] Link the Wiki Track, *INitiative for the Evaluation of XML retrieval (INEX)*, 2007. http://inex.is.informatik.uni-duisburg .de/2007/linkwiki.html.

[10] Kolbitsch, J. and Maurer, H. (2006) Community building around encyclopedic knowledge, *Journal of Computing and Information Technology*, 14.

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

[12] Adafre, S. F. and de Rijke, M. (2005) Discovering missing links in Wikipedia, *In Proceedings of the SIGIR 2005 Workshop on Link Discovery: Issues, Approaches and Applications*, Chicago, IL, USA, 21-24 August 2005.

[13] Jenkins, N. (2007) *Can We Link It*, http://en.wikipedia.org /wiki/User:Nickj/Can_We_ Link_It.

[14] Krowne, A (2003) An Architecture for Collaborative Math and Science Digital Libraries, *Thesis for Master of Science Virginia Polytechnic Institute and State University,* 19 July 2003.

[15] Gardner, J., Krowne, A. and Xiong, L. (2006) NNexus: Towards an Automatic Linker for a Massively-Distributed Collaborative Corpus, *In Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing,* 17-20 November 2006, 1-3.

[16] Bellomi, F. and Bonato, R. (2005) Network Analysis for Wikipedia, *In Proceedings of the 1st International Wikipedia Conference (Wikimania'05)*, Frankfurt am Main, Germany, 4-8 August 2005.

[17] Norris, J. R. (1997) Markov chains, *Cambridge Series in Statistical and Probabilistic Mathematics*, 2, Cambridge: Cambridge University Press.

[18] Ollivier Y. and Senellart P. (2007) Finding Related Pages Using Green Measures: An Illustration with Wikipedia, *In Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI'07)*, Vancouver, Canada, 22-26 July 2007.

[19] Schönhofen P. (2006) Identifying decument topics using the Wikipedia category network, *In Proceedings of the 2006 IEEE/EIC/ACM International Conference on Web Intelligence (WI'06)*, Hong Kong, 18-22 December 2006.

[20] Strube, M. and Ponzetto, S. P. (2006) WikiRelate! Computing Semantic Relatedness Using Wikipedia, *In Proceedings of the 21th National Conference on Artificial Intelligence (AAAI'06)*, Boston, Massachusetts, USA, July 2006, 16-20.

[21] Gabrilovich, E. and Markovitch, S. (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, *In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India, 6-12 January 2007.

[22] Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999) Trawling the Web for emerging cyber-communities, *Computer Networks*, 31(11-16), 1481-1493.

[23] Jeh, G. and Widom, J. (2002) SimRank: a measure of structural-context similarity, *In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)*, Edmonton, Canada, 23-26 July 2002, 538-543.

[24] Kessler, M. M. (1963) Bibliographic coupling between scientific papers, *American Documentation*, 14(10-25), 1963.

[25] Dean, J. and Henzinger, M. R. (1999) Finding related pages in the World Wide Web, *Computer Networks*, 1999, 31(11-16), 1467-1479.

[26] Kleinberg, J. (1998) Authoritative sources in a hyperlinked environment, *In Proceedings of the 9th Annual ACM–SIAM Symposium on Discrete Algorithms*, San Francisco, CA, USA, 25-27 January 1998, 668-677.

[27] Schamber, L. (1994) Relevance and Information Behavior, *Annual review of information science and technology 29,* Medford, NJ: Information Today, 3-48.

[28] Vakkari, P. and Hakala, N. (2000) Changes in relevance criteria and problem stages in task performance, *Journal of Documentation*, 56(5), 540-562.

[29] Kekäläinen, J. and Järvelin, K. (2002) Using graded relevance assessments in IR evaluation, *Journal of the American Society for Information Science and Technology*, November 2002, 53(13), 1120-1129.

[30] Wikipedia, the free encyclopedia, 2007. http://wikipedia.org/.

[31] Huang, W. C., Xu, Y., Trotman, A. and Geva, S. (2008) Overview of INEX 2007 Link the Wiki track, In Proceeding of INEX 2008 Workshop, Dagstuhl, Germany.

[32] Itakura, K. Y. and Clarke, C. L. A. (2007) University of Waterloo at INEX2007: Ad Hoc and Link-the-Wiki Tracks, *In Pre-proceedings of the INEX 2007 Conference,* Dagstuhl, Germany, 380-387.

[33] Reid, J., Lalmas, M., Finesilver, K. and Hertzum, M. (2005) Best Entry Points for Structured Document Retrieval – Part II: Types, Usage and Effectiveness, *Information Processing & Management*, 42(1), 89-105.

[34] D. Jenkinson, A. Trotman, Wikipedia Ad Hoc Passage Retrieval and Wikipedia Document Linking, *In Pre-proceedings of the INEX 2007 Conference,* Dagstuhl, Germany, 365-379.

[35] K. N. Fachry, J. Kamps, M. Koolen, J. Zhang, The University of Amsterdam at INEX 2007 *In Pre-proceedings of the INEX 2007 Conference,* Dagstuhl, Germany, 388-402.