Collaborative Knowledge Management: Evaluation of Automated Link Discovery in the Wikipedia

Wei Che 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

Using the Wikipedia as a corpus, the Link-the-Wiki track, launched by INEX in 2007, aims at producing a standard procedure and metrics for the evaluation of (automated) link discovery at different element levels. In this paper, we describe the preliminary procedure for the assessment, including the topic selection, submission, pooling and evaluation. Related techniques are also presented such as the proposed DTD, submission format, XML element retrieval and the concept of Best Entry Points (BEPs). Due to the task required by LTW, it represents a considerable evaluation challenge. We propose a preliminary procedure of assessment for this stage of the LTW and also discuss the further issues for improvement. Finally, an efficiency measurement is introduced for investigation since the LTW task involves two studies: the selection of document elements that represent the topic of request and the nomination of associated links that can access different levels of the XML document.

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, Evaluation, DTD, Best Entry Point

1.INTRODUCTION

The Wikipedia is a well-known online collaborative knowledge sharing system, a free encyclopedia that can be extended by any wiki contributor and modified by other wiki users [24]. At the time of writing, there are more than 75,000 active contributors working on more than 5,300,000 articles in more than 100 languages [25]. The growth in English Wikipedia articles had been around 100% per year from 2003 through most of 2006. There has been a close to linear increase in the number of articles since roughly September 2006, previous to that the trend was exponential.

Built upon traditional Wiki architectures, the functionality of the Wikipedia search engine is limited to title and full-text search. In general, it performs searching at the article level. After keywords have been entered in the search box, the ***Go*** function takes the user to the particular article while the ***Search*** function returns a list of ranked articles (including an estimate of relevance given as a percent) [26]. In addition to search facilities on the Wikipedia web site, there are a number of other search engines that search the encyclopedia (such as Google, Qwika, Lycos and Yahoo!).

Little research has been done in the area of semi-structured retrieval that can be directly applied to enhance the search features within the Wikipedia (although XML information retrieval studies have gained much attention in the last few years [19]).

Wikipedia contributors, like those of other Wikis must specify a variety of links that are relevant to a new article. They manually find and create links to other internal Wikipedia documents or external web pages. None the less, it is easy to find many un-related links that have been created and inserted in the documents (technical terms and years in particular). As an example, the term, *atomic transition probabilities*, in the *Albert Einstein* page had been split into *atomic transition* and *probabilities*, and *atomic transition* had been linked to the page *Transition rule*. However, in the list of search results for the term *atomic transition*, *Crystal field excitation* has the highest relevance (19.8%) and *Transition rule* second with relevance 12.3%. Similar to *atomic transition*, the term, *quantum theory*, had been linked to the article, *Quantum mechanics*, which is not found in the first page of results for *quantum theory*, but the *Quantum theory* page is returned as the most relevant result (100%).

By these examples, it is inappropriate to utilize standard search facilities to automatically nominate related links for anchor texts. A pilot track, Link-the-Wiki (LTW), launched by the Initiative for the Evaluation of XML Retrieval (INEX) in 2007 aims to provide a reusable resource and standard methods for the evaluation of automated link discovery within the English Wikipedia collection [8]. Previous work on link discovery exists of course (see section 8 for a brief review), but typically the methods operate on linking at the document level. As far as we know there has been no work published on automated discovery of document hyperlinks in the Wikipedia studying the choice of anchor texts and the link destination to specific positions *within* existing Wikipedia pages.

In this paper an assessment procedure for evaluating automated link discovery is proposed for use at INEX 2007 and beyond. In general, the procedure can be divided into several steps. First a number of orphan documents nominated by participants will be used as example link-less documents. Participants will generate links for these documents and submit results. Then the results will be pooled together for evaluation. Pooling will be performed manually. Finally, performance will be measured using agreed upon metrics. In this paper the pooling process will be discussed

SIGIR 2007 Workshop on Focused Retrieval
July 27, 2007, Amsterdam, The Netherlands

and future possibilities will be discussed. The challenges and the evaluation tool will also be introduced.

The remainder of this paper is organized as follows. In the next section, we survey the Wikipedia and its use as a corpus for focused information retrieval. Then we introduce the Link the Wiki track (in Section 3). Previous work that is related to the task of LTW is briefly summarized in section 4. In Section 5, we explain the terminology and present the submission format. In the next section (6), assessment steps and the evaluation process are introduced and discussed. Section 7 covers measures of search engine efficiency that will be considered for LTW track. The LTW 2007 track and its future scope are described in Section 8. Finally, conclusions and future work are provided in Section 9.

2.WIKIPEDIA AS AN IR COLLECTION

The Wikipedia is a free online document repository written collaboratively by wiki contributors around the world. Composed of millions of articles in numerous languages it offers many attractive features as a corpus for information retrieval tasks. In the first place, this wiki-based corpus is freely available so there are no distribution restrictions. The INEX Wikipedia collection has already been converted from its original wiki-markup text into XML [6]. That collection is composed of a set of XML files where each filename is a unique number corresponding to the id of the Wikipedia article (e.g. 16238.xml). Each file corresponds to an online article in Wikipedia (see Figure 1). A semantic annotation of the Wikipedia was also undertaken by others (e.g. [17]). Search as well as retrieval could benefit from rich semantic information in the XML Wikipedia collection, where it exists.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <article>
    <name id="588">Africa</name>
    <conversionwarning>0</conversionwarning>
  - <body>
      <templatePortal NAME="Portal" />
```

Figure 1. article XML format with corresponding id

In addition, the semi-structured format provided by the XML-based collection offers a useful property for the evaluation of various semi-structured retrieval techniques. Specifically, the linkage within a document is an especially interesting aspect of the Wikipedia and offers opportunities for investigating article categorization as well as the user interaction (e.g. browsing and searching) with a hyperlinked corpus.

The Wikipedia collection might be used for a variety of purposes such as XML information retrieval, machine learning, clustering, structure mapping, and categorization. The Wikipedia has already been used as an IR corpus in several evaluation initiatives. At INEX 2006 it was used for the evaluation of *ad hoc* XML retrieval and for the XML Document Mining track. At CLEF 2006, it was used as a corpus for question answering [23]. As the collection has already been used at INEX it is the natural choice for the INEX [8] 2007 Link-the-Wiki track.

3.LINK THE WIKI

For Link the Wiki at INEX 2007 the XML Wikipedia collection already used at INEX will be used as the document collection. It is composed of about 660,000 documents in English and is around 5GB is size. Many articles in the Wikipedia collection are already extensively hyperlinked.

The aim of the Link-the-Wiki task, first described by Geva and Trotman [21], is to offer an evaluation forum for proposing, testing, and discussing algorithms for evaluating the state of the art in automated link discovery in XML documents. The test collection including documents, judgments, and metrics for evaluating different systems and comparing various approaches to automated discovery of hypertext links will be made available for other researchers.

Participants will be given a set of (about 50) orphan Wikipedia documents nominated by participants. The task is two fold: first to analyze each orphan and to recommend anchor text and destinations within the Wikipedia; second to recommend incoming links from other Wikipedia documents. For 2007 we expect 25 anchor texts to be recommended for each orphan document. There will, therefore, be 1,250 outgoing links and 1,250 incoming links created for the 50 orphans, per submission. In future years the number of links might no longer be limited to 25 and links outside the Wikipedia (for example to the web) may be included.

Results will be submitted to the organizers who will pool them in the usual way. The pooled results will be analyzed and evaluated either automatically or by participants, depending on the kind of link. Article-to-article links can be evaluated by comparison with the original Wikipedia pages – they already contain relevant links created by the page authors. Links directly to XML elements must be evaluated manually.

The detailed experiment steps of assessment and evaluation will be described below in Section 6 and 7.

4.RELATED WORK

Since the goal of this paper is to propose the evaluation forum of Link-the-Wiki track as well as to significantly extend the tasks of link discovery to XML element level, we briefly introduce several instances of previous research on link analysis and generation, as well as document relevance identification, especially in the case of the Wikipedia. The past research described here is mainly targeted at the document level and the related evaluation for these approaches is manually performed.

While the Wikipedia has only gained much popularity in recent years, link analysis on the web and hypertext documents has been a relatively mature research field. Various link based techniques based on the correlation between the link (density) and the entities are analyzed and developed to deal with diverse research problems [1]. Links have been used to provide additional information for improving the quality of search engine results. Moreover, link analysis can also be used for topically classifying communities on the Web. The idea is to identify the implicit communities by the analysis of Web graph structure [13]. Kumar et al. also apply the concept of co-citation in the web graph for the similarity measure. Beside co-citation, bibliographic coupling and SimRank can be used to determine the similarity of objects (e.g. web pages), which are based on the citation patterns of documents and the similarity of structural context respectively [9][11]. Moreover, the Companion algorithm derived from HITS (**H**yperlink-**I**nduced **T**opic **S**election) is proposed for finding related pages by exploiting links and their order on a page [4][12].

This conducts a strategy of using a page's URL, instead of query terms, to search a set of related Web pages.

An overview of Wikipedia research was presented by Voss, which consists of different aspects of wiki studies [27]. This includes the visualization of wiki editing, relations of readers and authors, citation of wiki articles, the (hyperlinked) structure of Wikipedia and the statistic of Wikipedia. Recently, more research with regard to Wikipedia has been undertaken in particular for identifying the relevance of wiki articles. Bellomi and Bonato utilize network analysis algorithms such as HITS and PageRank to find out the potential relevance of wiki pages (content relevant entries) in order to explore the high level (hyperlinked) structure of Wikipedia and gain some insights about its content regarding to cultural biases [2]. Ollivier and Senellart have conducted a set of experiments for examining the performance of approaches on finding related pages within Wikipedia collection [14]. There are totally 5 methods included in the evaluation, including Green-based methods, *Green* and *SymGreen*, and three classical approaches, *PageRankOfLinks*, *Cosine* with tf-idf weight and *Co-citations*. The concept of these methods is to find out the most related neighborhood of a given node. They can be derived to achieve the task of finding the related pages.

Another interesting topic of utilizing an automated approach in finding related pages is to explore potential links in a wiki page. Adafre and de Rijke propose a method of discovering missing links in Wikipedia pages via clustering of topically related pages by LTRank and identification of link candidates by matching the anchor texts [1]. Jenkins presents a Wikipedia link suggestion tool, *Can We Link It*, for searching missing links in a page [10]. This suggestion tool can automatically eliminate those link candidates through the learning of user rejection and grammatical structure. However, some of these suggested links are still without merit with respect to the topic.

Furthermore, Wikipedia's category structures also offer useful information for topic identification. Schönhofen utilizes only the titles and categories of Wikipedia articles to characterize documents [18]. However, this simple method has not fully exploited the potential of Wikipedia, such as the internal text of articles, the category hierarchy and the linking structure of Wikipedia. *Wikirelate* proposed by Strube and Ponzetto uses Wikipedia to compute semantic relatedness of words through existing measures: Path based, Information content based and Text overlap based measures [20]. These measures mainly rely on either the texts of the articles or the category hierarchy. According to the shortcomings of *Wikirelate*, Gabrilovich and Markovitch 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 [6].

It is difficult to compare and contrast various approaches without a standard benchmark. This is the intent of the LTW track, while tightening and extending linking requirements to include BEPs.

5.TERMINOLOGY

Since the Link-the-Wiki track at INEX 2007 involves a series of new schemes and procedures, in this section, we will describe these in some details.

5.1Anchor Text Specification

Text file inversion is probably the most widely used technique in text retrieval systems [7]. For each term in an XML document a list of occurrences is maintained. The representation of each occurrence of a term is composed of the article id and term position within the XML document. We use this general representation in the specification of anchor text in the Link-the-Wiki task. Each term, phrase (or word gram) in an XML document can be located by identifying three parts: the filename (or article id in our case), the absolute XPath to the element in which the term is found, and the term or phrase position within the element.

The filename is used to identify the document within the XML collection. In the XML Wikipedia collection, a document file is presented by a unique id. For instance:

C:/Wikipedia/xml/23816.xml

The filenames are unique hence "**23816.xml**" is sufficient to unambiguously identify the document.

An XML element within the document may be identified by the absolute XPath expression relative to the file's root element (see Table 1).

Table 1. The absolute XPath expression

Absolute XPath Context
/article[1]/body[1]/section[5]/section[2]/p[4]
/article[1]/body[1]/p[1]/emph2[1]
/article[1]/body[1]/section[4]/item[3]/collectionlink[3]

Finally, with the XML document object model (DOM) it is possible to specify a particular text node character position. In the following expression the last number is the term position that identifies the start position (in characters) of the term within the specific XPath context:

/article[1]/body[1]/section[2]/p[1]/text()[6].3

In the Link-the-Wiki task we are proposing to identify anchor text start and end character positions in this manner.

5.2Example Specification of Link Discovery

With the element specification format described above, the LTW task can accept submissions that work with anchor text and links to specific XML elements. We use the term *best entry point* (**BEP**) as already used in INEX to describe a destination element within a document from which to start reading. Anchor text must be identified precisely by using the DOM as it is a passage of text and not an XML element or a simple location within the text.

An example submission is depicted in Figure 2. As shown each topic (orphan page) is identified by a topic-id, file name, and title. While these attributes are the same for each topic, and are thus interchangeable, all three are included for the sake of convenience and clarity. For each orphan two sets of links are identified - *outgoing* and *incoming*.

Outgoing links are composed of a set of *link*s from the orphan page to existing Wikipedia pages. Each *link* consists of an

anchor and a target *file* and a *best entry point* within that file. Collectively these identify a unique XML element in a Wikipedia document.

Incoming links are composed of a set of *link*s from anchor texts within existing Wikipedia pages to a *best entry point* in the orphan page.

To work with document to document (e.g. "see-also") links all that is required is the specification of all XPath expressions as **/article[1]**. In this case the entire topic specification is degenerates to a set of links between documents without any explicit anchor or best entry points. This is a deliberate decision made to accommodate low-cost entry into the Link-the-Wiki track.

```
<topic id="38" file="13876.xml" name="Albert Einstein">
  <outgoing>
    <link>
      <anchor>
        <start> /article[1]/body[1]/p[3]/text()[2].10 </start>
        <end> /article[1]/body[1]/p[3]/text()[2].35 </end>
      </anchor>
      <linkto>
        <file> 123456.xml </file>
        <bep> /article[1]/sec[3]/p[8] <bep>
      </linkto>
    </link>
    …
  </outgoing>
  <incoming>
    <link>
      <anchor>
        <file> 654321.xml </file>
        <start> /article[1]/body[1]/p[3]/text()[2].10 </start>
        <end> /article[1]/body[1]/p[3]/text()[2].35 </end>
      </anchor>
      <linkto>
        <bep> /article[1]/sec[3]/p[8] <bep>
      </linkto>
    </link>
    …
  </incoming>
</topic>
```

Figure 2. Sample submission

5.3DTD

A Document Type Definition (DTD) will be defined for specifying the XML document structure. It will contain a list of legal elements and attributes from within the Wikipedia collection. This will allow participants to validate their runs before being submitted. Although, since document-to-document linking will be the default at INEX 2007, this is not immediately needed. A full version of LTW will be run in future years so the full DTD will be needed at that stage. The DTD for LTW 2007 is depicted in figure 3.

```
<!ELEMENT topic (outgoing, incoming)>
<!--
          topics DTD
                                    -->
<!ELEMENT outgoing (link+)>
<!ELEMENT incoming (link+)>

<!ATTLIST topic
           id CDATA #REQUIRED
           file CDATA #REQUIRED
           name CDATA #REQUIRED>
<!--
          The links
                                    -->
  <!ELEMENT link (anchor,linkto)>

  <!ELEMENT anchor (file?, start, end)>
  <!ELEMENT linkto (file?, bep)>

  <!ELEMENT file  (#PCDATA)>
  <!ELEMENT start  (#PCDATA)>
  <!ELEMENT end  (#PCDATA)>
  <!ELEMENT bep  (#PCDATA)>
```

Figure 3 The LTW assessment DTD

5.4Specific XML Elements

The XML data model offers extensible element tags which can be arbitrarily nested in order to capture semantics [4]. Information such as titles, references, sections and sub-sections are explicitly captured using nested, application specific XML tags.

The use of XML elements as the retrieval unit is believed to provide a more accurate result than using whole documents. But, as yet using XML structure has not proven useful in XML *ad hoc* retrieval [22], except for some very specific queries such as multimedia queries that specifically target images. None the less, it is the XML-IR functionality that is required for Link-the-Wiki. Links from automatically identified anchor-text to best entry points in a document are needed.

The evaluation of the Link the Wiki task will require a different and possibly more complicated method of evaluation then XML-IR. Link evaluation is very different from conventional precision / recall so far used to evaluate XML-IR at INEX. Specifically a score is needed for the identification of anchor-texts as well as for the corresponding best entry point destinations. Although standard INEX metrics such as BEPD (see [8]) might be used for the latter, scoring the former remains unaddressed.

5.5Best Entry Points (BEPs)

At INEX a best entry point (BEP) is a specific document element from which the user can perform some optimal access to a series of relevant document elements [15]. The purpose of a BEP is to complement the users' searching activities and facilitate direct entry to relevant items within documents. The identification of BEP is already a sub-task in the *ad hoc* track at INEX and the

methods that are used there may be used in Link-the-Wiki essentially unmodified.

The BEP results in the LTW submission can be expressed in the following format.

<bep> /article[1]/sec[4]/p[3]</bep>

6.EVALUATION METHODOLOGY

The Link-the-Wiki track will be held once a year and will be generally based on the following steps:

1. Participants nominate 10 or more topics (Wikipedia pages) of reasonable length for which link discovery will be performed. These pages must (obviously) exist in the XML Wikipedia collection.
2. Topics are distributed to participants who run their link discovery search engines. A submission for each topic consists of a list of selected anchor texts and corresponding links to best entry points within the Wikipedia collection. The number of incoming and outgoing links will be restricted to some reasonable and manageable number for each topic, perhaps based on topic length.

   In the initial year (2007) participants will specify links at any level of granularity, but evaluation will only be performed between whole articles (all links will be treated as "see-also" links). The Wikipedia currently contains only this kind of link and not links to best entry points.
3. The *pooling* process is performed to merge results from different participants and that correspond to the same document. The specific details of this are tied to the design of an assessment tool and are outside the scope of this paper. The pooled results for see-also links can be automatically assessed by comparison to links in existing documents. In future years it will be necessary for assessors to manually assess links.
4. The link discovery search engines will be scored with respect to performance using standard metrics (that are yet to be defined). We expect and encourage experimentation with several metrics since the best way to score runs is not immediately obvious.
5. The results are returned to participants who in turn analyze, present, and discuss their approaches at the INEX workshop.

The detailed processes will be described in the following sections.

6.1Procedure

An initial set of LTW topics are nominated by participants and a final set of at least 50 topics will be selected. These topics (Wikipedia documents) will then be orphaned by eliminating the anchor texts and their associated destinations (the XML tags, *collectionlink*, will all be discarded). The *"what links here"* information will also be discarded from the topic documents. A topic submission should identify no more than 25 anchor texts from any part of the document and identify the most relevant destinations. Furthermore, no more than 25 incoming links can additionally be identified.

In 2007 submission of BEP links will not be required, but rather document to document links will suffice. However, anchor texts should be specified. Ideally, the link engines of participants should be able to automatically find the 25 most relevant anchor texts in response to the content of given topics and specify the associated link at the XML element level in the INEX documents. This means that clicking the anchor text does not lead to an article but to the particular document Best Entry Points.

At INEX 2007 evaluation will be performed between articles only so submissions may contain BEPs but they will be automatically reduced to whole articles in evvaluation. In future years, and with the use of an assessment tool, the evaluation of more precise link specifications will be supported.

For use with an assessment tool the pooling process will need to execute once the results are all submitted. Each nominated topic will then be associated with a set of links for assessment. The pooled results might be assessed in one of two ways: automated assessment might be performed by comparing results in the pool with those already in the orphan to get a *precision* and *recall* score. Manual assessment might be used to individually assess links. The exact details of the metric and the assessment tool are outside the scope of this paper and are yet to be defined in precise detail. This will be done through discussion between track participants.

6.2Challenges

The preliminary procedure of assessment has been stated and described above. However, the detailed methodology (e.g. approaches and metrics) are still not finalized. In fact, much like it was with the *ad hoc* track at INEX, one would expect that only after some considerable experimentation with evaluating LTW submissions could a methodologically sound evaluation approach be put in place.

In terms of different element levels, article level evaluation is not dissimilar to standard *ad hoc* retrieval and some form of F-Score might be utilized. Given an orphan document, taking into account the *precision* and *recall* of identified links (both incoming and outgoing), computing some form of mean may be sufficient. The hypothesis is that a relative comparison of runs will be sufficient to derive an appropriate ranking score [16].

With automated evaluation, there is no exhaustive assessment. Consequently, some returned links may be appropriate, but not already appear in the Wikipedia. The consequence is that evaluation results may appear pessimistic.

Manual assessment is expected to be more accurate, but is time consuming. With a suitable assessment tool we believe that this effort can be reduced to reasonable levels. The design of an efficient assessment tool is currently underway.

With exhaustive assessment pooling becomes important. Pooling with the LTW is more problematic than with a traditional Cranfield experiment since there is a real possibility that there will be very little overlap between submissions. In particular, the anchor texts from the runs may only partially overlap, or not overlap at all, and links may be pointing to different BEPs. This can lead to unreliable evaluation as observed when traditional *ad hoc* pools are too shallow.

At present we are exploring ways to collect the entire set of links from all submissions, eliminate duplicates where possible, and assess all remaining links. This will at least ensure that evaluation of the systems that contribute to the pool is meaningful. It is neither clear how re-usable such a set of assessments will be nor how exhaustive a set of manually assessed links can be. This can only be studied after the track has produced the first set of results.

Figure 4. Link the Wiki Submission Interface

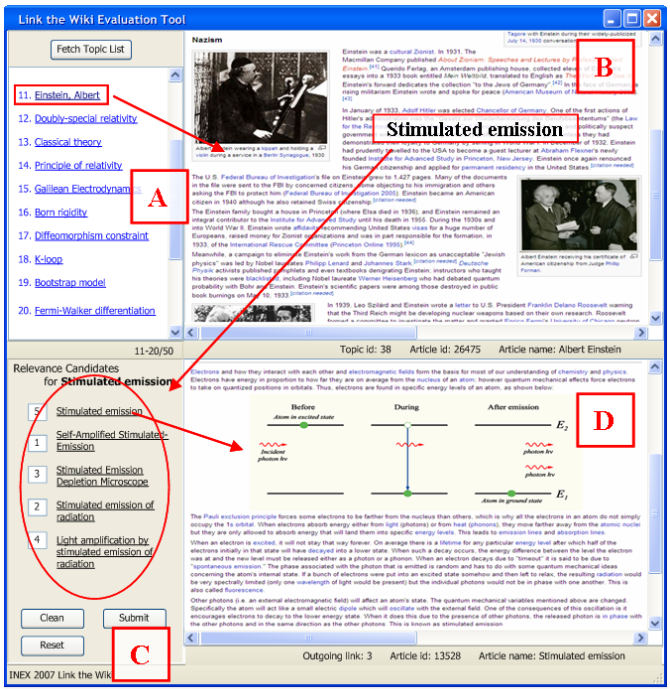

Figure 5. Link the Wiki evaluation tool

6.3Tools

An (online) assessment system will be provided for the LTW community for various evaluation scenarios. The preliminary prototype is illustrated in Figure 5.

In section A, a list of topics is displayed. The topic content is shown on the right hand side in section B. Once the user clicks on the anchor text in the topic content, a set of candidates associated with the anchor will be given in section C. A selected link in section C will show the corresponding linked-to text in section D. In this manner a user can see both the anchor text in context, and the linked-to text in context. A text box will be used to enter a relevance score for the selected link. The user can navigate through different links by clicking on link names.

This tool can also be used to view submissions as well as the pool. Section C displays the associated links with the rsv (score) from the participants' system while the content of a link is shown in section D. All anchor texts (or elements) that link to this content will be highlighted in the document in section B. This interface provides an easy way for participants to examine their result sets as well as to navigate through different anchor texts and linked contents.

7.EFFICIENCY

Missing from INEX has been any measure of search engine efficiency. Although the precision of the *ad hoc* runs is measured each year, how long it took the search engine to produce the results is completely unknown, participants don't normally publish this detail. The Link-the-Wiki track will be the first track at INEX in which efficiency will be considered.

Ideally each participant will run their solutions on the same computer configuration; however this is not feasible for several reasons: first, it is not practical to prescribe a given computer and operating system configuration and to expect participants to build it; second, prolonged use of such a machine will inevitably result in changes to the configuration (for example operating system patches might in some cases be installed but not in others); third, shipping a machine between participants is costly and time consuming and will result in changes to the configuration as an increasing number of search engines are installed; finally, bringing the search engines to the machine (for example at the workshop) is also not possible as search engines may not be portable across operating systems and doing so might start and operating system battle.

For these reasons participants will be asked to submit their runs and to state (as part of their run) the time it took for their system to produce the set of results. All this will be defined in the run submission DTD. Participants will also be asked to include configuration of the machine on which the run was generated.

It might appear at first inspection that the problem is that of building the optimal implementation of the optimal solution and running it on the fastest computer available. However, optimality is hard to define and there is a time/performance trade-off. For link discovery this is of particular interest.

An optimal set of links could be identified by a human with complete knowledge of the document collection – however it would be costly to gain such knowledge and to employ such an individual. An immediate set of results might be gained by building a finite state automaton from the titles of all documents

in the collection and a simple parser. A better (but more time consuming) result might be found with a part of speech (POS) tagger and some natural language processing. For a semi-commercial entity wanting to build an open source repository such as the Wikipedia, different subscription levels might be offered depending on the cost (in CPU cycles) of the quality of the linking.

For the purpose of comparison, using different algorithms on different machine configurations leads to some problems: like is not being compared with like and the group with the fastest machine should be able to produce the fastest runs. It is not yet clear how to address this problem and it is expected as a topic of debate. One solution is to also solicit from participants an estimate of the dollar cost of the machine (or machines) on which the result was generated. With a time and a cost, the unit of measure might be the precision-dollar. This, again, is likely to cause debate as the price of a machine increases exponentially with performance and the measure would favor network of workstation (NOW) or pile of PC (POPC) configurations - but perhaps justly.

A real-time question answering exercise was conducted at CLEF 2006. We believe that that exercise and the efficiency testing of Link-the-Wiki are the beginning of a new era in forum evaluation that started with the TREC Web Track. Once the limits of precision begin to be approached, small mutually-exclusive sets improvements begin to proliferate and are of less interest than substantial cost reductions in producing the results. This is already being seen in full-document retrieval where techniques such as impact-ordering and index pruning have been proposed. For *ad hoc* XML-IR no such techniques have yet been proposed or tested. We anticipate the *ad hoc* track at INEX adopting an efficiency task and consequently fast and effective XML-IR search engines. It should however be noted that the Link-the-Wiki task in itself will demand significantly more processing per topic than an *ad hoc* topic. Relatively few topics will mote likely suffice to severely tax slow underlying IR systems.

8.LINK-THE-WIKI 2007 AND BEYOND

In 2007 participants will submit 10 orphans each. A set of at least 50 will be distributed. For each one, each participant will identify 25 anchor-texts and for each anchor-text 5 best entry points. The time it took to generate the result and an approximate US dollar cost for the hardware will also be submitted. Runs will be reduced to a set of document-to-document links and performance measured with yet to be announced metrics.

In future years, metrics that score both anchor-text identification and the best entry point identification will be added. Links to destinations outside the Wikipedia are likely to be added too.

9.CONCLUSIONS AND FUTURE WORK

The Wikipedia is an attractive corpus for performing automated link discovery experiments since the collection is extensively hyperlinked. In this paper, we briefly described the objective and requirements of Link-the-Wiki track as well as the preliminary procedure of assessment and evaluation. The task of LTW has gone beyond traditional information retrieval that normally searched relevant whole article. The Link-the-Wiki task represents, in our view, an ideal use case for XML-IR. It aims at accessing different element levels within an XML document, which presents the most relevant components (sections, paragraphs, etc.) in relation to anchor text selected from the topic of request.

Briefly, the process of assessment can be depicted as follows. At least 50 orphan pages will be given to participants for LTW tasks. The automated discovery of document hyperlinks at the different XML element levels is performed by the participants' systems. The results are submitted to the organizers. The submissions are analyzed and the elements as well as the associated links on each topic are examined and selected as the candidates for the final evaluation. At the first stage, 25 anchor texts for each topic with the related 5 destinations will be chosen for manual evaluation. Since this pooling process and the final evaluation are manual and time-consuming, the automated approaches and the standard metrics will be investigated further first, especially for the element level evaluation (e.g. anchor text to BEP).

In addition to the evaluation, there are many options for improving the work introduced in this paper. Although the precision of the results for both the selection of elements that represent the topic and the retrieval of links associated with the elements is important, the efficiency measure is another consideration in the real world retrieval systems. Response time, the time a user must wait for a result, considers the CPU and I/O latency. An efficient LTW system will certainly be an asset to the Wikipedia and other collaborative knowledge management systems.

10.REFERENCES

[1] Adafre, S. F. and de Rijke, M. 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.

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

[3] Chernov, S., Iofciu, T., Nejdl, W. and Zhou, X. Extracting Semantic Relationships between Wikipedia Categories, *In First Workshop on Semantic Wikis: From Wiki to Semantic [SemWiki2006]*, Budva, Montenrego, 12 June, 2006.

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

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

[6] Gabrilovich, E. and Markovitch, S. 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.

[7] Geva, S. and Leo-Spork, Murray XPath Inverted File for Information Retrieval, *In proceedings of INEX 2003 Workshop*, Schloss Dagstuhl, 15-17 December 2003, 1-8.

[8] INitiative for the Evaluation of XML retrieval (INEX), 2007. http://inex.is.informatik.uni-duisburg.de/2007/

[9] Jeh, G. and Widom, J. 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.

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

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

[12] Kleinberg, J. 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.

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

[14] Ollivier Y. and Senellart P. 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.

[15] Reid, J., Lalmas, M., Finesilver, K. and Hertzum, M. Best Entry Points for Structured Document Retrieval – Part II: Types, Usage and Effectiveness, *Information Processing & Management*, 42(1):89-105, 2006.

[16] Salem, M., Woodley, A. and Geva, S. IR of XML documents? A Collective Ranking Strategy, In Proceeding s of Third International Workshop of the Initiative for the Evaluation of XML Retrieval, Dagstuhl Castle, Germany, 113-126.

[17] Schenkel, R., Suchanek, F. M. and Kasneci, G. YAWN: A Semantically Annotated Wikipedia XML Corpus, *In 12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, Aachen, Germany, 2007, 277-291.

[18] Schönhofen P. Identifying document 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.

[19] Sigurbjornsson, B., Kamps, J. and de Rijke, M. Focused Access to Wikipedia, *In Proceedings of the sixth Dutch-Belgian Information Retrieval workshop (DIR 2006)*, TNO ICT, Delft, The Netherlands, 13-14 March, 2006.

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

[21] Trotman, A. and Geva, S. Passage Retrieval and other XML-Retrieval Tasks, *In Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, Seattle, Washington, USA, 10 August 2006, 48-50.

[22] Trotman, A. and Lalmas, M. Why Structural Hints in Queries do not Help XML-Retrieval, *In Proceedings of the 29th Annual International ACM SIGIR Conference*, Seattle, Washington, USA, 6-11 August 2006, 711-712.

[23] WiQA: Question answering using Wikipedia, 2006. http://ilps.science.uva.nl/WiQA/index.html

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

[25] Wikipedia: Size_of_Wikipedia, 2007. http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

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

[27] Voss, J. Measuring Wikipedia, *In Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2005)*, Stockholm, Sweden, 24-28 July 2005.