The Methodology of Manual Assessment in the Evaluation of Link Discovery

Wei Che (Darren) Huang
Faculty of Science and Technology
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 Science and Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

Abstract - *The link graph extracted from the Wikipedia has often been used as the ground truth for measuring the performance of automated link discovery systems. Extensive manual assessments experiments at INEX 2008 recently showed that this is unsound and that manual assessment is essential. This paper describes the methodology for link discovery evaluation which was developed for use in the INEX 2009 Link-the-Wiki track. In this approach both manual and automatic assessment sets are generated and runs are evaluated using both. The approach offers a more reliable evaluation of link discovery methods than just automatic assessment. A new evaluation measure for focused link discovery is also introduced.*

Keywords

Wikipedia, Link Quality, Manual Assessment, Evaluation.

1. Introduction

The Wikipedia free encyclopedia is the most popular collaborative information repository on the web. It continues to enjoy increasing popularity amongst web users as well as amongst a diverse set of knowledge content editors [1]. Wikipedia documents are densely linked in the traditional way, from text anchors in one document to a target document. Although external links to other web pages outside the Wikipedia also exist, the link structure within the Wikipedia is quite different from that of the Web. The use of hyperlinks on the Web tends to vary, ranging from elaboration to referential to navigational. Text anchors do not necessarily denote the concept of the target, and even if they do they often take the user to a different but related web site.

The Wikipedia link structure is typically built by matching text anchors to semantically related entries. Most links within the Wikipedia have a strong semantic relationship between the anchor context and the target context. The purpose of a Wikipedia link is almost invariably to provide more detailed information about something. The majority of the links are conceptual rather than navigational.

In a growing collection, such as the Wikipedia, the maintenance of the link graph can quickly become more time-consuming and complicated than adding content. Newly created documents should be linked to from text anchors in existing pages. Links to deleted documents must be erased. There is also general maintenance of the link graph for documents that change or are extended.

Several [2, 3, 4, 5] automated *link discovery* algorithms have been proposed as methods to alleviate the link maintenance problem. The INEX *Link-the-Wiki* Track [6] takes the traditional link discovery problem a step further with *focused link discovery*. The aim is to identify text anchors in a source document and a best entry point (*BEP*) in a target document. In HTML a BEP is a named anchor and an anchor-to-BEP link is specified using *#name* on the end of the target document's URL.

Focused systems are potentially more useful to the user because of the reduced need to navigate (especially in a long document or on a mobile device). Anchor-to-BEP link discovery is also a harder (and more interesting) problem than *anchor to document* discovery because of the focused relationship between the anchor context and the target document BEP context. The current method of link discovery is based on the page name matching or similarity. A broad range of technologies, e.g. natural language processing, data mining, machine learning, information retrieval, information extraction and link discovery, are encouraged to integrate to resolve the issue of linking anchor to best entry points.

After two years of INEX experiments it appeared as though the problem of the file-to-file link discovery was solved. Two fundamentally different approaches (anchor link analysis [3] and page name analysis [2]) could identify high quality links when evaluated against links already in the Wikipedia. Near perfect precision scores at high recall levels were seen.

However, after extensive manual assessment of INEX runs it became clear that the use of the existing link graph lead to biased and optimistic evaluation [7]. It appears as though the near perfect scores are achieved because a substantial proportion of the links in the Wikipedia are in fact generated automatically using similar methods to those used by the link dis-

Proceedings of the 14th Australian Document Computing Symposium, Sydney, Australia, 4 December 2009.

covery systems being evaluated. Manual assessment appears to be essential for robust evaluation of link quality.

There are other reasons for manually assessing link discovery systems at INEX:

1. There appear to be many links in the Wikipedia that are not useful. Some links are inserted automatically and may not be considered relevant by users of the Wikipedia (for instance, links to *year* documents).
2. The Wikipedia is largely linked from an anchor to a whole document; best entry points are rarely seen. It is, therefore, not possible to use the existing Wikipedia link graph to evaluate *anchor to BEP* link discovery systems.
3. In the Wikipedia it is quite reasonable to expect some anchors to target multiple destinations. There could, for example, be a variety of thematic links, multilingual links, or links which extend the anchor's context with varying degrees of complexity (simple vs. full Wikipedia). The existing link graph does not support the evaluation of systems which support multiple links per anchor discovery.

The need for a robust and standardized manual assessment and evaluation methodology is the motivation for this paper. We hope that this methodology will be adopted for link discovery experiments beyond INEX 2009.

2. Wikipedia

There are more than 200 different language versions of the Wikipedia (September 2009). They are freely available as a database and are particularly well suited to IR experiments.

Between 2006 and 2008 INEX used a dump of the English Wikipedia consisting of 659,388 documents. For 2009 INEX has used a fresh dump consisting of 2,666,192 documents. The documents were converted from the original Wiki-markup to XML.

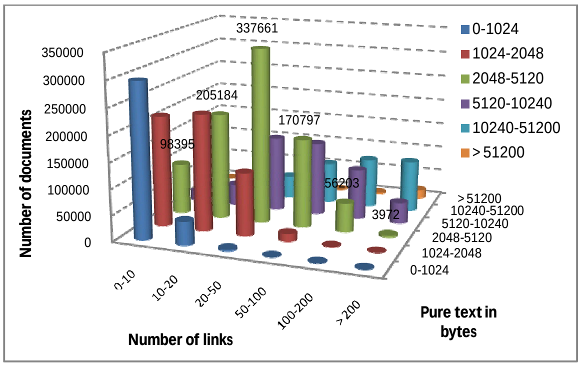


Figure 1: Relationship between document length and number of link in the INEX 2009 Wikipedia collection.

Presented in Figure 1 is the relationship between document size and the number of outgoing links. There are no very short documents with a high link count and there are very few long documents with few links. Most documents link to between 10 and 100 different pages within the collection.

There are 24,168 homonym disambiguation pages. These pages are not suitable for link discovery experiments as they are (essentially) content-free.

3. Related Work

The 2008 INEX Wikipedia collection [8] was converted from Wiki-markup into XML for XML-IR experiments. The 2009 INEX collection was also converted into XML but for use in a broader set of experiments. One point of difference between the two collections is the semantic annotations present in the 2009 collection (see Schenkel et al. [9]).

The Wikipedia has already been used as an IR corpus in several evaluation initiatives. Since INEX 2006 it has been used for the evaluation of ad hoc XML retrieval and for XML Document Mining. At CLEF 2006, it was used for question answering [10].

A link suggestion tool, *Can We Link It*, was developed by Jenkins [11]. It extracts a number of anchors which have not been discovered in an article and that might be linked to other Wikipedia documents**.** Using this tool the user can add new anchors and corresponding links back from a Wikipedia article. Mihalcea & Csomai present the *Wikify* [4] system. It integrates automatic keyword extraction and word sense disambiguation to identify the important concepts in a document and links these to corresponding documents in the Wikipedia.

Link discovery systems are typically evaluated against the Wikipedia itself. Pages are selected as IR topics, the algorithms are run over the topics, and the result compared to the links that are already in the document. Mihalcea & Csomai [4] used the Turing test to further validate their results. Milne & Witten [5] used the Mechanical Turk to solicit links for the AQUAINT collection. INEX considers link discovery to be a recommender task and so the results list is ranked; set based evaluation is inappropriate.

Two evaluation frameworks, DIRECT [12] at CLEF and EPAN [13] at NTCIR, provide a GUI and modules for evaluation. INEX assesses all topics, and also uses a GUI evaluation tool for ad hoc retrieval.

4. Experimental Methodology

A subset of the Wikipedia collection is chosen as a topic set. All anchor links to and from the topics, from and to the collection, are removed (orphaning the documents). Specifically, a random set of (6600 in 2008, 5000 in 2009) documents was chosen as the topics for file-to-file linking; track participants nominated topics (50 in 2008, 33 in 2009) for anchor-to-BEP linking. The goal is to identify both outgoing and incoming links from and to those topics.

INEX offers two linking tasks: file-to-file and anchor-to-BEP. The former is a low-cost entry-level task for new participants (and as a sanity check for the latter task). The task is to identify up to 250 documents that the topic should link to; no anchor or BEP need be identified.

In the anchor-to-BEP task the system can identify up to 50 outgoing anchor texts per topic. For each anchor at most 5 target document/BEP pairs are allowed. For incoming link discovery, a set of at most 250 anchors (in the collection) targeting BEPs in the topic are to be identified. Both incoming and outgoing links are from anchor to BEP.

A text anchor is identified by its position (offset and length) within the document. A BEP is identified by its position. Positions are specified as character offsets (excluding markup) from the document start.

Participants were invited to submit runs. In total 30 runs were submitted in 2008. It was prior to the 2009 submission deadline at the time of writing.

5. Manual Assessment

5.1 Methodology

In 2008 two sets of assessments were generated, one from the Wikipedia and the other from the runs.

The Wikipedia ground-truth assessment set consisted of just those links already in the Wikipedia. It is an automatically generated set of links from anchors to documents.

Submission runs were *pooled*. The pooling process combines overlapping anchor texts to form a pool-anchor which is presented to the assessor. A pool-anchor might contain a number of anchors as well as a set of target BEP links. All the links already in the Wikipedia topic were added to the pool. The pool was then manually assessed.

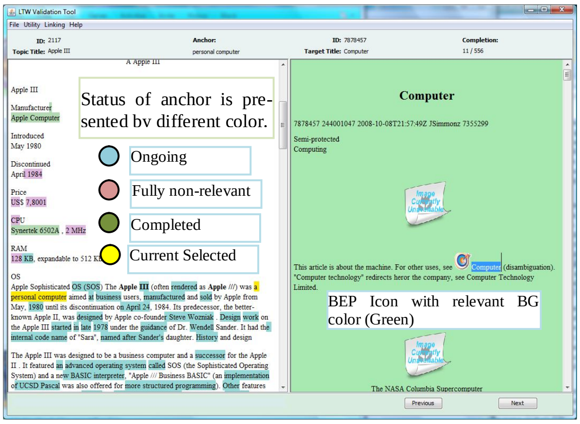
For the purpose of evaluation it is assumed that all non-assessed links are non-relevant. However, as the pool was exhaustively assessed, there is a reusability issue and does not affect submitted runs. We note that the same convention is used in other forums and tracks (such as TREC).

A validation tool was provided and distributed to assist developers of focused link discovery systems. It allowed participants to view their submissions in an interface similar to that used by the assessors. The tool helped participants debug their submissions (the calculation of BEP can be non-trivial), as well as perform sanity checks on their algorithms.

5.2 Assessment

Built on experience using the INEX ad hoc assessment tool, a GUI-based relevance assessment tool (i.e. *GPXrai*) was custom designed and built for the manual assessment of link discovery pools (see Figure 2). The interface is comprised of a split screen.

The topic pane is located on the left hand side. The right hand pane is used to show the target document. Two distinct assessment modes are provided, one for outgoing links, the other for incoming links.



ID: 2117
Topic Title: Apple III
Anchor: personal computer
ID: 7878457
Target Title: Computer
Completion: 11 / 556

A Apple III

Apple III

Manufacturer
Apple Computer

Introduced
May 1980

Discontinued
April 1984

Price
US$ 7,8001

CPU
Synertek 6502A , 2 MHz

RAM
128 KB, expandable to 512 KB

OS
Apple Sophisticated OS (SOS) The **Apple III** (often rendered as **Apple** ///) was a personal computer aimed at business users, manufactured and sold by Apple from May, 1980 until its discontinuation on April 24, 1984. Its predecessor, the better-known Apple II, was designed by Apple co-founder Steve Wozniak . Design work on the Apple III started in late 1978 under the guidance of Dr. Wendell Sander. It had the internal code name of "Sara", named after Sander's daughter. History and design

The Apple III was designed to be a business computer and a successor for the Apple II . It featured an advanced operating system called SOS (the Sophisticated Operating System) and a new BASIC interpreter, "Apple /// Business BASIC" (an implementation of UCSD Pascal was also offered for more structured programming). Other features

Status of anchor is presented bv different color.

● Ongoing
● Fully non-relevant
● Completed
● Current Selected

Computer

7878457 244001047 2008-10-08T21:57:49Z JSimmonz 7355299

Semi-protected
Computing

This article is about the machine. For other uses, see Computer (disambiguation). "Computer technology" redirects heror the company, see Computer Technology Limited.

BEP Icon with relevant BG color (Green)

The NASA Columbia Supercomputer

Figure 2: INEX 2009 Link-the-Wiki Assessment Tool

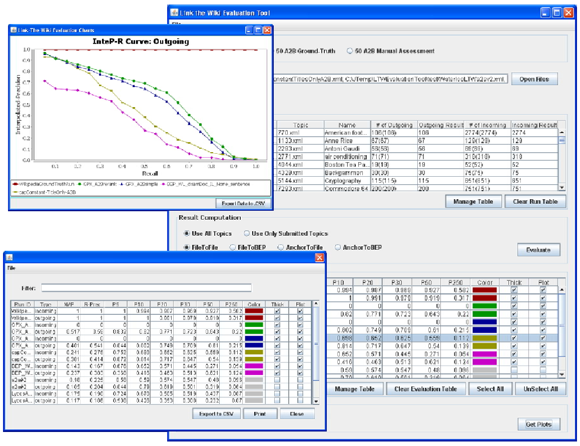
Outgoing links are initially assessed. The topic document is displayed with highlighted pool anchors. In the first instance the assessor goes through all the anchors and rejects (a mouse right-click) those which are obviously irrelevant. The pool can contain many such anchors, for instance year or other coincidental links. Each link for each remaining anchor is then displayed, in turn, with the right pane showing the target document. The assessor then either rejects (a mouse right-click), or accepts the target as relevant (by double-clicking to indicate the BEP, then mouse left-click).

Incoming links are assessed in a similar manner, but the locations of the anchor and BEP are swapped. Now the anchors are from other documents and the BEPs are inside the topic document. The assessor is required to accept or reject each prospective link.

In INEX 2008 the pools contained between 405 and 1722 links. Assessment logs suggest that between 4 and 6 hours were required to assess a topic. On average, only 7.4% of a pool was judged relevant.

6. Evaluation

A portable (Java) evaluation tool, *LtwEval*, was developed for evaluation purposes. It is GUI based and provides numerous evaluation metrics including: precision, recall, MAP, and precision@R. Different runs can be evaluated and compared to each other. Precision/recall graphs can be generated for sets of runs (see Figure 3). Anchor-to-BEP runs can be evaluated as either file-to-file or anchor-to-BEP. The tool was distributed to participants.



...onstantTitlesOnlyA2B.xml; C:\JTemp\LTWEvaluationTool\test\WaterlooLTWa2av2.xml;

Topic	Name	# of Outgoing	Outgoing Result	# of Incoming	Incoming Result
770.xml	American foot...	106(106)	106	2774(2774)	2774
1133.xml	Anne Rice	67(67)	67	129(129)	129
2293.xml	Antoni Gaudi	56(56)	56	69(69)	69
2771.xml	air conditioning	71(71)	71	310(310)	310
4044.xml	Boston Tea Pa...	19(19)	19	52(52)	52
4329.xml	Backgammon	30(30)	30	75(75)	75
5144.xml	Cryptography	115(115)	115	851(851)	851
7293.xml	Commodore 64	200(200)	200	751(751)	751

Result Computation

● Use All Topics ○ Use Only Submitted Topics

● FileToFile ○ FileToBEP ○ AnchorToFile ○ AnchorToBEP

Run ID	Type	MAP	R-Prec	P5	P10	P20	P30	P50	P250	Color	Thick	Plot
Wikipe...	incoming	1	1	1	0.994	0.987	0.969	0.927	0.582	🟥	☑	☑
Wikipe...	outgoing	1	1	1	1	0.991	0.979	0.919	0.317	🟥	☑	☑
GPX_A...	incoming	0	0	0	0	0	0	0	0	🟩	☑	☑
GPX_A...	outgoing	0.517	0.59	0.832	0.82	0.771	0.723	0.643	0.22	🟩	☑	☑
GPX_A...	incoming	0	0	0	0	0	0	0	0	🟦	☑	☑
GPX_A...	outgoing	0.401	0.541	0.844	0.802	0.749	0.709	0.61	0.215	🟦	☑	☑
capCo...	incoming	0.241	0.275	0.752	0.698	0.652	0.625	0.559	0.112	🟨	☑	☑
capCo...	outgoing	0.381	0.414	0.872	0.814	0.717	0.647	0.54	0.139	🟨	☑	☑
BEP_W...	incoming	0.143	0.167	0.676	0.652	0.571	0.445	0.271	0.054	🟪	☑	☑
BEP_W...	outgoing	0.237	0.303	0.358	0.416	0.463	0.513	0.621	0.124	🟪	☑	☑
s2a#2	incoming	0.16	0.225	0.56	0.59	0.574	0.547	0.48	0.056	⬜	☐	☐
s2a#2	outgoing	0.165	0.204	0.644	0.79	0.619	0.501	0.319	0.064	⬜	☐	☐
LycosA...	incoming	0.175	0.196	0.724	0.670	0.505	0.519	0.437	0.067	⬜	☐	☐
LycosA...	outgoing	0.117	0.166	0.506	0.426	0.353	0.308	0.232	0.07	⬜	☐	☐

Figure 3: INEX 2008 Link-the-Wiki Evaluation Tool

The "best" metric to use for focused link discovery evaluation is not obvious. As with all metrics, it is important to first define the use-case of the application. The assumption at INEX is that link-discovery is a recommendation tasks. The system produces a ranked list of anchors and for each a set of recommended target/BEP pairs. The user navigates a limited number of anchors and selects only a few to embed in the new document.

A link discovery system might identify a very large number of possible links. The Wikipedia has a page for each letter in the Latin alphabet and so each letter of each word might be linked. It also contains potentially overlapping links, for example there is a page for *world*, a page for *war*, and a page for *world war*. The user is expecting the system to identify relevant anchors and links, and to place these at the top of the results list. The list should also be comprehensive because it is not clear that the document author can know a priori which links will be relevant to a reader of the document. That is, link discovery is a recall oriented task

The Mean Average Precision based metrics are very good at taking rank into account and are recall oriented. They are also very well understood. A good metric for link discovery should, consequently, be based on MAP. The difficulty is computing the relevance of a single result in the results list.

For the anchor *The Theory of Relativity*, an equally good anchor might be *Relativity*. For evaluation purposes it is assumed that if the target is relevant and the anchor overlaps a relevant anchor then the anchor is relevant; $f_{anchor}(i) = 1$. This is subtly different from the *world war* problem above, different in so far as the target must also be relevant. Of course, this definition of relevant anchor aids in reusability.

The assessor might have assessed any number of documents as relevant to the given anchor. If the target of the anchor is in the list of relevant document then it is considered relevant; $f_{doc}(i) = 1$. In the INEX ad hoc track the BEP is considered to be subjective. If the search engine can put the relevant passage on the user's screen then it is considered a "hit". The contribution of the links' BEP is a function of distance from the assessor's BEP:

$$f_{bep}(j) = \begin{cases} \dfrac{n - 0.9 \times d(x,b)}{n} & if\ 0 \le d(x,b) \le n \\ 0.1 & if\ d(x,b) > n \end{cases}$$

Where $d(x,b)$ is the distance between submission BEP and result BEP in character. Therefore, the score of $f_{bep}(j)$ varies between 0.1 (i.e. d is greater than n) and 1 (i.e. the submission and result BEPs are exactly matched). The score of 0.1 is reserved for the right target document with an indicated BEP not in range of n. n typically is set up as 1000 (characters). The score of a result in the results list is then:

$$P = \left[(f_{anchor}(i)) \times \frac{\left(\sum_{j=1}^{m} \left(f_{doc}^{i}(j) \times f_{bep}^{i}(j)\right)\right)}{m_i} \right]$$

Where m is the number of returned links for the anchor and m_i is the number of relevant links for the anchor in the assessments. As the result list is restricted to 5 targets per anchor m_i is capped at 5 for evaluation. A perfect run can thus score a MAP of 1.

7. Conclusion and Outlook

Although it has appeared as though link discovery is a solved problem, manual assessment of participants runs at INEX 2008 showed that, in fact, it is not. The INEX result raises new questions about methodologies for link discovery evaluation, and in particular focused link discovery systems.

In this contribution we propose and describe a new comprehensive methodology. This methodology is based on manual assessment of link relevance. A new metric is proposed to measure the performance of a run. Our methodology is being used for the INEX 2009 Link-the-Wiki track.

Our further work will focus on evaluation quality and on the efficiency of the manual assessment. This will be done using assessor surveys and interviews.

We remain fascinated by the appalling performance of the Wikipedia itself when evaluated against the manual assessments. It is our expectation that, once the methodology is stable, link discovery systems will outperform human created hypertext links.

References

[1] Alexa, The Web Information Company *http://www.alexa.com/topsites.*

[2] Geva, S., *GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia*, INEX 2007, pp. 404-416.

[3] Jenkinson, D., K.-C. Leung, A. Trotman, *Wikisearching and Wikilinking*, INEX 2008, pp. 374-388.

[4] Mihalcea, R., A. Csomai, *Wikify!: linking documents to encyclopedic knowledge*, CIKM 2007, pp. 233-242.

[5] Milne, D., I.H. Witten, *Learning to link with wikipedia*, CIKM 2008, pp. 509-518.

[6] INEX (2009) http://www.inex.otago.ac.nz/tracks/wiki-link/wiki-link.asp.

[7] W.C. Huang, Trotman, A., Geva, S (2009), *The Importance of Manual Assessment in Link Discovery*, SIGIR 2009, pp. 698-699.

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

[9] Schenkel, R., Suchanek, F. M., Kasneci, G. (2007) YAWN: *A Semantically Annotated Wikipedia XML Corpus*, BTW 2007, pp. 277-291.

[10] WiQA: Question answering using Wikipedia (2006) http://ilps.science.uva.nl/WiQA/index. html

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

[12] Mitanura, T., Nyberg, E. Shima, H., Kato, T., Mori, T., Lin, C.Y., Song, R., Lin, C. J., Sakai, T., Ji, D., Kando, N., *Overview of the NTCIR-7 ACLIA Taska: Advanced Cross-Language Information Access* NTCIR-7, pp. 11-25.

[13] Di Nunzio, G. M. and Ferro, N., *DIRECT: A System for Evaluating Information Access Components of Digital Libraries*, ECDL 2005, pp. 483-484.