The Importance of Manual Assessment in Link Discovery

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

Using a ground truth extracted from the Wikipedia, and a ground truth created through manual assessment, we show that the apparent performance advantage seen in machine learning approaches to link discovery are an artifact of trivial links that are actively rejected by manual assessors.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: System and Software – *Information networks, Performance evaluation (efficiency and effectiveness).*

General Terms

Documentation, Performance, Experimentation, Human Factors

Keywords

Wikipedia, Link Discovery, Assessment, Evaluation, INEX.

1. INTRODUCTION

Maintenance of hypertext links between documents in a centralized document repository is problematic for several reasons: when new documents are added old documents must be updated to point to the new; when old documents are deleted all links to the deleted document must be removed; if the topical content of a document changes over time then links must be added, deleted, and updated. In a growing collection, such as the Wikipedia, the maintenance can quickly become more time-consuming than adding new content. This maintenance requirement was motivation for the INEX Link-the-Wiki track, a standard evaluation forum for automated link discovery in a closed document repository.

The track methodology proceeds as follows: Take a snapshot of the Wikipedia. From that snapshot, extract one document and eradicate all links to and from that document from and to the collection (orphan the document). Using the orphan as an IR topic, identify a ranked list of links to (and from) that document into (and out of) the collection. Repeat the process a large number of times. Finally, measure the performance of the link discovery system against the ground-truth as is in the pre-orphaned documents. Different from the work of Milne & Witten [5] and of Mihalcea & Csomai [4], a *ranked* list of links is required because INEX considers link detection systems to be recommender systems and as such assumes a human will read a list of results.

After two years of the track it appeared as though identifying high quality outgoing links was solved. Jenkinson *et al.* [3] submitted a run based on the work of Itakura & Clarke [2] and of Geva [1] which scored a mean average precision (MAP) score of 0.73. The run maintained high precision even at moderately late points of recall (for example, a precision of 0.85 at a recall of 0.5). High precision and recall scores for non-ranked lists were also seen by Milne & Witten but, as they identify, a direct comparison of prior work in this field is not possible because different versions of the Wikipedia and different topics have been used.

The relative ease of achieving high scores motivates our research question. We ask whether, or not, identifying links similar to those already present in a Wikipedia document is a task helpful to users. Superficially it is obvious that it is, as the ground truth to which the comparison is made is the human edited Wikipedia itself. Mihalcea & Csomai conducted a Turing test of their system generated pages against the pre-orphans and showed that the two are "hardly distinguishable" while Milne & Witten used the Mechanical Turk to evaluate linking documents in the AQUAINT corpus and show similar performance to automatic assessment against a ground truth extracted from the pre-orphans.

We show that comparing to a ground-truth extracted from the Wikipedia is unsound. We do this by comparing the performance of link discovery systems on the same orphans against two assessment sets: one extracted from the pre-orphans; the other is a manually assessed superset of this that also includes all links identified by runs submitted to the INEX 2008 Link-the-Wiki track.

2. METHODS

The INEX Wikipedia collection consisting of 659,388 documents was used for the experiments. Each of the 10 groups participating in the track was asked to nominate, from the collection, 5 documents to be orphaned for the experiment. 50 topics were nominated and all nominated topics were used in the experiment.

Links within the collection to and from the pre-orphans were extracted and used as the ground truth to which runs were compared. This formed the AUTOMATIC set. Mihalcea & Csomai and Milne & Witten use such a set for training and testing the performance of their algorithms before additionally manually validating.

Using standard methodology, the orphans were sent to participating groups, each group ran their link discovery system and returned a ranked list of at most 50 outgoing text anchors (each of which targeted up-to 5 documents) for each orphan. The results were pooled and the AUTOMATIC set was added to the pool. Pools were manually assessed to completion, by the group that nominated the orphan. This formed the MANUAL set.

Assessment against the AUTOMATIC set provides a score for the performance of a run relative to the Wikipedia. Assessment against the MANUAL set provides for the additional scoring of links identified in a run, but not present in the Wikipedia. As it is convention to assume all non-assessed links to be non-relevant, the larger MANUAL set can catch cases where a run contains links not in the AUTOMATIC set (and so considered irrelevant) but that are relevant. We, consequently, expect a comparison

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

against the two sets to result in higher performance against the MANUAL set – assuming all AUTOMATIC links are relevant.

3. MANUAL ASSESSMENT

In total 30 runs were submitted, pools contained between 405 and 1722 links. Figure 1 shows the software especially designed for assessment; on the right is the pool, left the orphan, and middle is the link target document. Assessors selected links and then marked anchors, targets, or both as relevant or not. We estimate that between 4 and 6 hours was spend assessing each topic. On average 7.4% of a pool was judged relevant.

4. RUNS

Two fundamentally different approaches to link discovery are seen in the INEX runs, our analysis is on one of each approach:

Anchor link analysis is due to Itakura & Clarke [2]. First, all anchor-texts and target documents used in the collection are identified. Next, the document frequencies of the anchor-text in the whole collection are identified. Finally, the anchor-texts from the collection are identified in the orphan and the most probable target document chosen. Links are ranked on ratio of the target document frequency to anchor text document-frequency. We use the corrected[1] Jenkinson *et al*. [3] run, *Otago*, for this investigation. The uncorrected run ranked 1st at INEX 2008.

Page name analysis is due to Geva [1]. If a document title is ever seen in the orphan then a link to the document is added. Preference is given to longer over shorter titles. We use the corrected Geva run, *QUT*. The uncorrected run ranked 13th at INEX 2008.

A third run, *Wikipedia*, was artificially generated directly from the orphan documents by using just the *first* 50 links seen in the pre-orphaned documents. This is the best possible run.

5. RESULTS

The precision recall graph in Figure 2 shows the performance of the three runs against the AUTOMATIC assessments. As expected, run *Wikipedia* scores perfectly, *Otago* performs well, and *QUT* performs adequately. In Figure 3 the same three runs are assessed using the MANUAL set. It can be seen there that the runs are tightly clustered whereas using AUTOMATIC assessment they are not. It can also be seen that the run derived from the Wikipedia performs little-better than the other two. A two-tailed *t*-test shows all runs significantly differ (at 1%) from each other using AUTOMATIC assessment but no run is significantly different from any other (even at 5%) using MANUAL assessment.

6. DISCUSSION AND CONCLUSIONS

Run *Otago* performs very well against the AUTOMATIC assessment set, but Trotman[2] recently showed the performance of that run is limited by the 50-targets track restriction. Near perfect scores can be achieved using *anchor link analysis*. Shown here, however, is that the same run does not perform as well when assessed against the MANUAL set – and neither does run *Wikipedia*! There are several reasons why this might be observed:

Some of the hypertext links present in the Wikipedia documents are trivial (such as dates). When the human assessors in the experiment were presented with these links they usually actively rejected them – the Wikipedia contains many non-relevant links.

Disagreement between existing Wikipedia links and the human assessor is expected. In future work we will examine this further, and measure its effect on the relative rank order of INEX runs.

The implications of our finding are twofold: contrary to prior results, using links present in the Wikipedia in the *anchor link analysis* approach (i.e. machine learning) produces results little-better than simply just choosing document titles; but more importantly, measuring the performance of link discovery relative to a ground truth extracted from the Wikipedia is unsound.

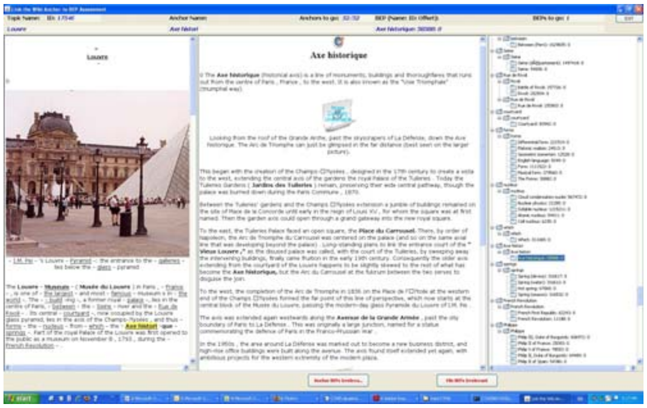

Figure 1: INEX 2008 assessment tool.

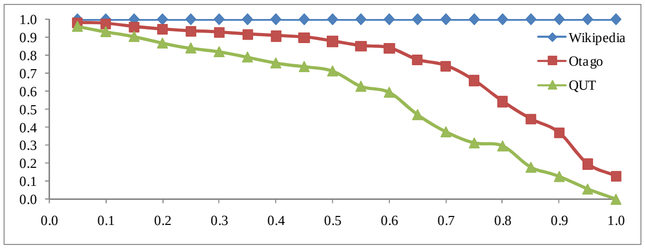

Figure 2: AUTOMATIC assessment (Precision/Recall)

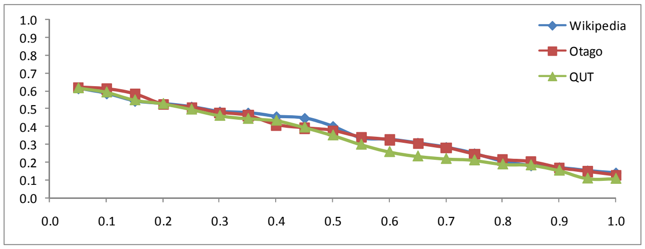

Figure 3: MANUAL assessment ((Precision/Recall)

REFERENCES

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

[2] Itakura, K.Y., C.L. Clarke, *University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks*, INEX 2007.

[3] Jenkinson, D., K.-C. Leung, and A. Trotman, *Wikisearching and Wikilinking*, pre-proceedings of INEX 2008.

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

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

[1] Corrected after INEX but before the published proceedings.

[2] Unpublished, but in his presentation at INEX 2008.