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

Wei Che Huang[1], Shlomo Geva[1] & Andrew Trotman[2]

[1]Faculty of IT
Queensland University of Technology (QUT)
Brisbane, Australia

[2]Department of Computer Science
University of Otago
Dunedin, New Zealand

Presentation layout

- Background and Motivation
- Related Work
- Link-the-Wiki (LTW @ INEX2007)
  - Setup
  - Assessment
  - Evaluation
- Extensions beyond 2007

Background and Motivation

From the Wikipedia "Wiki" page:

"A wiki is a collaborative website which can be directly edited by anyone with access to it. Ward Cunningham, developer of the first wiki WikiWikiWeb, originally described it as "the simplest online database that could possibly work". Wikipedia is one of the best known wikis."

… Collaborative Knowledge Management

- The Wikipedia is probably the most popular Wiki
- Entries are typically contributed by individual users
- Entries are then maintained collectively
- Extensive hyper linking is supported –
  - User initiated
  - Automatically suggested links
  - Link bots

… Collaborative Knowledge Management

- Hyperlinks within the Wikipedia can be very useful for ranking algorithms
- Third-party search engines “crawl” the Wikipedia and present results therein with high rank
- Pre-computed hyperlinks can be effectively used by search engine crawlers
- The more extensive and accurate the collection’s link structure is, the more effective subsequent search by a third party can be.

… Collaborative Knowledge Management

Making a new entry:

- Links from the new entry require good knowledge of the existing collection
- Existing entries may benefit from links to the new entry
- Links at the document level may not be as useful as links from Anchor to to Best Entry Point (BEP)

LTW vs Ad-Hoc IR

- In Ad-Hoc IR the query itself is typically short and devoid of user context
- Document matching methods are often restricted to the input query (e.g. standard Google queries)
- To incorporate context –
  - Blind feedback can be used to try and infer context from preliminary results (can be fragile)
  - Explicit/implicit relevance feedback can be obtained, at the cost of additional user interaction, in query refinement
  - Explicit context information can be obtained through more complex and direct user interaction

LTW vs Ad-Hoc IR …

- In the LTW task the initial “query” is typically short too (anchor text)
- The anchor text’s neighbourhood provides considerably more context (e.g. the embedding passage or XML element)
- By extending the neighbourhood, context can be expanded from passage or element to the entire document

LTW vs Ad-Hoc IR …

- The LTW task operates on whole documents and links from Anchor to BEP rather than Anchor to document - in fact, links could be to within the same document
- LTW is perfectly suited to XML collections which provide a standard protocol for defining anchors, hyper links, and targets
- The computational demands on LTW topics are greater – by orders of magnitude (a topic is an entire orphan document requiring numerous links)
- Performance evaluation is different (more later…)
- LTW facilitates collaborative curating, beyond searching

Related work

- Link analysis is not new. It predates the WWW and even IR as an established discipline.
- Work on link discovery within the Wikipedia is not new (see paper for references) but is typically aiming for document level links
- Automated link bots and link suggestion tools – for the Wikipedia - already exist http://can-we-link-it.nickj.org/ (Jenkins)

Related work

Apparently missing from early work:

- Anchor to BEP links
- Evaluation of LTW

Link-the-Wiki (LTW @ INEX2007)

- An evaluation forum and a set of standard tasks and corresponding achievable results.
- A reusable resource for evaluating and comparing different state of the art systems and approaches to automated link discovery.
- More specifically, given a new orphan wikipedia document, the task is to analyse the text and recommend a set of incoming and outgoing links from/to anchor text in the existing collection.
- Going beyond traditional text document analysis, in the context of INEX we aim to operate at the anchor text to BEP level using an XML collection and relying on XML standards.

Sample submission

```
<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[5]/text()[3].15 </start>
        <end> /article[1]/body[1]/p[5]/text()[3].25 </end>
      </anchor>
      <linkto>
        <bep> /article[1]/sec[2]/p[3] <bep>
      </linkto>
    </link>
      ...
  </incoming>
</topic>
```

Evaluation Methodology

- Select N topics from existing Wikipedia articles
- Orphan the documents
- Participants run LTW search engines on topics and submit results
- Results are pooled
- Links are assessed (how?)
- Runs are evaluated (how?)

Topic Selection

- In 2007 we have asked each participant to nominate 10 topics
- Topics received from 11 groups and about 90 selected (some were missing from the INEX collection, others changed a lot)

Orphan the documents

- Collection links are eliminated:
  - <collectionlink>
  - <Wikipedialink>
  - <Languagelink>
  - <Redirectlink>
  - <Unknownlink>
- External links are left intact:
  - <Outsidelink>
  - <Weblink>

Orphan the documents…

- Eliminate the topic files from the collection
- Eliminate all references to the topics from other files in the collection
- How to orphan?
  - physically (redistribute the collection)
  - Virtually (“ignore” topics and references) – 2007

Link assessment

- In 2007
  - document level assessment only
  - Using the existing links as the ground truth
- Problems –
  - the existing links may not be sufficiently exhaustive
  - the existing links may have been incorrectly generated by link bots without adequate scrutiny

Link assessment…

- 2008 onwards
  - full Anchor to BEP link assessment
  - a distributed assessment tool deployed

Evaluation

- Adopt conventional approaches as much as possible -
  - Precision and Recall of links
  - Link BEP score
  - Link anchor text score
- Evaluate runs over all links
- Evaluate separately incoming links and outgoing links

Evaluation 2007

- To kick start the track –
  - Evaluate links at the article level
  - Use the existing collection links as the ground truth
  - No manual assessment
  - Generous limits on run lengths – a large number of links can be assessed automatically

Evaluation beyond 2007

- An assessment tool is required
- Distributed tool rather than centralised server based tool
- Assessment scoring granularity has to be resolved
  - Binary (accept/reject link)
  - Graded
- Evaluation software for linking at the anchor->BEP level

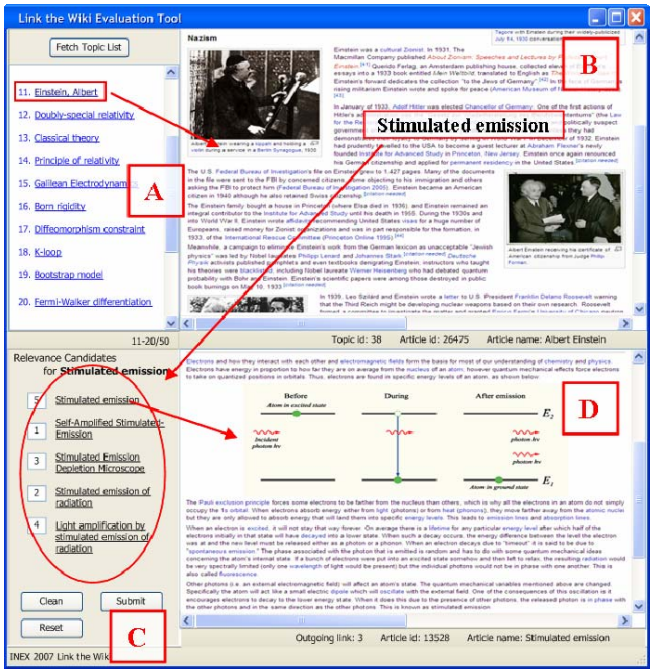

Figure 5. Link the Wiki evaluation tool