OUC's participation in the 2012 INEX Book and Linked-Data Tracks

Michael Preminger[1], Ragnar Nordlie[1], David Massey[1], and Nils Pharo[1]

Oslo and Akershus University College of Applied Science

Abstract. In this article we describe the Oslo University College's participation in the INEX 2012 endeavor. This year we participate in the Book Track's "Prove it" (as in 2011) and Social search tasks, as well as the Linked Data track's Ad-hoc task.
In 2011, the OUC submitted retrieval results for the "Prove It" task with traditional relevance detection combined with detection of confrmation based on specificity detected through the Wordnet concept hierarchy. In line with our belief that proving or refuting facts are different semantic aware actions of speech, we have this year attempted to incorporate some semantic support based on Named entity recognition.
For the Social search task, we wish to examine the utility of the MARC-data (subject heading field) in social searching for readings.
For the Linked data task, we wish to explore the possibility of using links as a query expansion mechanism.

1 The Prove-it task of the Book Track

In recent years large organizations like national libraries, as well as multinational organizations like Microsoft and Google have been investing labor, time and money in digitizing books. Beyond the preservation aspects of such digitization endeavors, they call on finding ways to exploit the newly available materials, and an important aspect of exploitation is book and passage retrieval.

The INEX Book Track[1], which has been running since 2007, is an effort aiming to develop methods for retrieval in digitized books. One important aspect here is to test the limits of traditional methods of retrieval, designed for retrieval within "documents" (such as news-wire), when applied to digitized books. One wishes to compare these methods to book-specific retrieval methods.

One important mission of such retrieval is supporting the generation of new knowledge based on existing knowledge. The generation of new knowledge is closely related to access to – as well as faith in – existing knowledge. One important component of the latter is claims about facts. This year's "Prove It" task may be seen as challenging the most fundamental aspect of generating new knowledge, namely the establishment (or refutal) of factual claims encountered during research.

On the surface, this may be seen as simple retrieval, but proving a fact is more than finding relevant documents. This type of retrieval requires from a passage to "make a statement about" rather than "be relevant to" a claim, which traditional retrieval is about. The questions we posed in 2010 were:

- what is the difference between simply being relevant to a claim and expressing support for a claim
- how do we modify traditional retrieval to reveal support or refutal of a claim?

We also made the claim that "Prove It" sorts within the (not very well-defined) category "semantic-aware retrieval", which, for the time being will be defined by us as retrieval that goes beyond simple string matching, and is aware of the meaning (semantics) of text.

Those question, being rhetorical in part, may be augmented by the questions

- How can one detect the meaning of texts (words, sentences and passages) and incorporate those in the retrieval process to attain semantic-aware retrieval

and consequently

- can one exploit technologies developed within the semantic web to improve semantic-aware retrieval

The latter is not directly addressed in this paper, but we claim that the techniques used here point in this direction.

1.1 Task Definition and User Scenario

The prove-it task is still at its infancy, and may be subject to some modifications in the future. Quoting the user scenario as formulated by the organizers

> The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to confirm or refute a given factual statement. Users expect to be pointed directly at book pages that can help them to confirm or refute the claim of the topic. Users are assumed to view the ranked list of retrieved book pages starting from the top of the list and moving down, examining each result. No browsing is considered (only the returned book pages are viewed by users).

This user scenario is a natural point of departure as it is in the tradition of information retrieval and facilitates the development of the task by using existing knowledge. As a future strategy, it may be argued that this user scenario is gradually modified, as ranking in the context of proving is a highly complex process, and, in the context where Prove-it algorithms are most likely to be used, arguably superfluous.

1.2 What Is a Proof?

What constitutes a proof is well defined in fields like mathematics and computer science. In connection with a claim or a statement of fact, it is less obvious what demands a passage of text should satisfy in order to be considered proof of the claim. Obviously, we are looking for a passage which expresses a relevant truth

about the claim, but what are the characteristics which signal a sufficient degree of relevance and truthfulness? We might want to identify a trustworthy passage, which in turn might be identified by considering the source of the passage, the degree to which the passage agreed with other passages treating the same claim or fact, or the centrality of the claim to the main content of the text. We might want to identify a concentrated passage, a passage where the largest amount of elements contained in the claim were represented or where they were by some measure most heavily represented. We might look for a definitional passage, which typographically or linguistically showed the characteristics of a definition. Or we might try to identify a "proof" by linguistic characteristics, mostly semantic, which might be of different kinds: certain typical words might be relatively consistently used to speak about a fact or claim in a "proving" manner, writing in a "proving" mode might entail using terms on a certain level of specificity, etc. These latter aspects are orthogonal to the statement or claim itself in the sense that they (at least ideally) apply equally to whatever claim being the subject of proving / confirming.

1.3 Semantic Approaches to Proof

A statement considered as a "proof" (or confirmation) may be characterized semantically by several indicators:

- the phenomenon to be supported may be introduced or denoted by specific terms, for instance verbs indicating a definition: "is", "constitutes", "comprises" etc.
- terms describing the phenomenon may belong to a specific semantic category
- nouns describing the phenomenon may be on a certain level of specificity
- named entities of different kinds are heavily used
- verbs describing the phenomenon may denote a certain type of action or state

Deciding which specificity level or which semantic categories will depend on the semantic content and the relationship between the terms of the original claim. Without recourse to the necessary semantic analysis, we assume that in general, terms indicating a proof / confirmation will be on a relatively high level of specificity. It will in some way constitute a treatment of one or more aspects of the claim at a certain level of detail, which we expect to be reflected in the terminology which is applied.

In 2011, we were investigating whether terms, in our case nouns, found on a page indicated as a potential source of proof diverges in a significant way from other text in terms of level of specificity. We determined the level of noun specificity through their place in the WordNet([2]) term hierarchies.

In this year's experiments, we proceed along the same line, trying to detect named entities. This years effort represents a starting point in taking into use named entity detection to assist in identifying confirming pages. Confirmation or proofs will often be about subjects identifiable by a name. Gradually, we first

need to find the limits of current detection of named entities, how easy it is to adapt it to a relatively diverse text mass that the (English part of) our text collection is, and then the approach's effectiveness in detecting proving pages. The two main possibilities in taking NED into use are:

- Detecting of named entities in general: pages that mention many named entities are candidates for being "confirming of something". Other methods are used to find the specific subject of proof. this means we only detect named entities in the book pages.
- Detecting the named entity being the subject of the statement to be proved. This means detecting named entities in the query, and in the.

Even though the latter possibility looks obvious it entails some problems, like polymorphism in identification of entities, which must be approached. This is the main rationale for starting out with the former possibility.

1.4 Ranking According to "Proof Efficiency"?

In this paper we are still following the two-step strategy of first finding pages relevant to the claim, and from those pages trying to identify pages that are likely to prove the claim[1]. The first step is naturally done using current strategies for ranked retrieval. The second stage identifies *among relevant documents* those which prove / confirm the statement. Rank order is not necessarily preserved in this process: if document A comprises a better string-wise match with the claim than does document B, document B can still be more efficient at proving the claim than document A is. Not all elements that make a document relevant also make it a good prover

Another issue is the context in which prove-it is used. One example is the writing of a paper. A writer is (again, arguably) more likely to evaluate a greater number of sources for proof of a claim than he or she would in a context of pure fact finding. Additionally, different contexts would arguably invite different proof emphases. All this advocates for use of other strategies of presenting proving results than ranked lists.

1.5 Indexing and Retrieval Strategies

The point of departure of the strategies discussed here is that confirming or refuting a statement is a simple action of speech that does not require from the book (the context of the retrieved page) to be *about* the topic covering the fact. In this way the "Prove It" task is different than e.g. the one referred to in [3] This means that we do not need the index we build for search purposes to be context-faithful (pages need not be indexed in a relevant book context). It is the formulation of the statement in the book or page that matters.

[1] We see refutal as a totally different type of task and will not address it in this paper.

1.6 Indexing

In line with the above, indexing should facilitate two main aspects at retrieval time: identifying relevant pages and finding which of these is likely to prove a claim. The first aspect is catered for creating a simple index of all the words in the corpus, page by page. The pages are treated as separate documents regardless of the book in which they appear. The second aspect is catered for by

1.7 Named entity discovery

Named entity discovery is a natural language processing (NLP) activity. There exist several tools that perform NED. The choice this time fell on the opennlp package of the Apache project. The package was used with default settings (no special training), with the assumption that the big diversity of the book collection is not apt to any significant improvement with respect to the default settings.

1.8 Runs and Results

2 Social Book search

2.1 Introduction

The social book search features two representations of books: the social data, which is a mixture of "Amazon data" (descriptive and social data to facilitate book sale via Amazon) and social encounters as recorded in the libraryThing fora on one hand, and, on the other hand, traditional library data (MARC records) entered by professional catalogers. The main purpose is to find out the relative utility of each of these representations when it comes to automatic book recommendation.

[**?**] has attempted to compare the utility of social data to that of DEWEY classification data (which are also available in the Amazon records). In this paper we try to build upon this, and look at subject headings extracted from the MARC data.

2.2 Indexing and retrieval strategies

The collection has been loaded to a database where all types of data about each book are associated with the book's ISBN. We create an Indri index that includes all the tagged XML information that is extracted from both amazon and the LT fora. To each Indri document (book representation) we also add a section containing the subject headings extracted from the MARC record or records of that book[2].

[2] Some of the books contain MARC records from both the Library of congress as well as the British Library, with a greater diversity of subject headings.

<DOC>
<DOCNO>0525949283</DOCNO>
<TEXT>
<SH>Balloon ascensions Women balloonists</SH>
<bookdoc><book><isbn>0525949283</isbn><title>The Little Balloonist</title><ean>9780525949282</ean><binding>Hardcover</binding><label>Dutton Adult</label><listprice>$21.95</listprice><manufacturer>Dutton Adult</manufacturer><publisher>Dutton Adult</publisher><readinglevel/><releasedate/><publicationdate>2006-01-19</publicationdate><studio>Dutton Adult</studio><edition/><dewey>813.6</dewey><numberofpages>224</numberofpages><dimensions><height>70</height><width>580</width><length>850</length><weight>75</weight></dimensions><reviews><review><authorid>A1HA6KZZNDCME9</authorid><date>2007-02-25</date><summary>More like 2 1/2 stars...</summary><content>History collides with fiction in THE LITTLE BALLOONIST set ... may feel that the overall story is a bit rushed and that the possible depths that could have been conveyed just never emerged.

COURTESY OF CK2S KWIPS AND KRITIQUES</content><rating>2</rating><totalvotes>2</totalvotes><helpfulvotes>1</helpfulvotes></review><review><authorid>A1UDDVTG2K1K72</authorid><date>2006-02-07</date><summary>Ahh...A Wonderful Love Story for Valentine's Day</summary><content>What a love story! A perfect present for Valentine's Day if you're still looking for any last minute gifts. I bought this book on a friend's recommendation and read it from start to finish in one evening. A lost love? A rekindled romance? All definitely keep the pages turning. And Donn's vivid descriptions of Napoleonic France all come alive. I highly recommend.

</content><rating>5</rating><totalvotes>0</totalvotes><helpfulvotes>0</helpfulvotes></review></reviews></browseNode></browseNodes></book></bookdoc>
</TEXT>
</DOC>

Every possible element in this XML is made known to the indexing system, so it can be used as an extent e.g. for retrieval time weighting. An obvious strategy here is to weight <SH> at retrieval time when trying to find the effect of the subject headings

One problem we have is that only about two-thirds of our books have MARC records. This means that full comparison of the utility uses less documents. Still,

with many enough topics we may hope that a good number of those do have relevant *and* judged books among those with MARC records[3].

At retrieval time, weighting of subject headings can be done in the following way:

```
<query>
  <number>530</number>
<text>
   #combine ( #weight ( 1.0 #combine( Jesus.text Why.
       text scholarly.text perspective.text historical.
       text From.text ) 2.0 #combine( Jesus.sh Why.sh
        scholarly.sh perspective.sh historical.sh From.
       sh ) ) )
</text>
</query>
```

3 Preliminary runs and results

Preliminary runs were performed in accordance with the description in Section2.2. Table 1 summarizes the results. There seems to be a problem with the basic setting that makes it difficult to assess the contribution of the subject headings. This makes us refrain from further analysis at the present moment, apart from the suspicion that the results may be due to insensitive use of elements from the data collected from each book.

Run name	Query	Element weight	map	ndcg	ndcg_10	recip
sb_g0	Title only	all elements equal	0,0128	0,1713	0	0,0357
sb_2xh	Title only	sh (2) all (1)	0,02	0	0,0045	0,02
sb_g_ttl_nar	Title and narrative	all elements equal	0,065	0,1785	0,0756	0,1615
sb_g_ttl_nar_2xh	Title and narrative	sh (2) all (1)	0,03	0,016	0	0,03

Fig. 1. preliminary results for the Social search book track

4 The linked data track, the Ad-Hoc task

4.1 Introduction

The purpose of the linked data track is to find out how techniques within the semantic web / linked data can be used to improve and enhance retrieval of Wikipedia articles. The data collection is an XML'ified version of a Wikipedia

[3] We do not know it at the time of writing.

subset (about 4.1M articles), where incoming and outgoing links are tagged in terms of RDF-properties (DBpedia), and the article text is also included.

Our experiment is based on a two-stage approach. The initial search is in an index built from the entire corpus. Here each article is only represented by heading and category texts. For each topic, the initial search result (1000 articles) is enhanced by articles that form triples with the initial articles (as subjects or objects). This process results in a set of typically several thousands articles, including the initial set. Those are used to create a smaller "topic-wise index" with more data on each of the retrieved articles. In addition links in the articles are collected to enhance the results with the most popular articles that are not captured in the initial search.

4.2 Indexing and retrieval strategies

Stage one Prior to building the main index, a filter removes from the corpus files that are not considered to be articles. This included files that describe images. Files that had titles with the prefixes 'File:', 'Wikipedia:', 'Category', 'Portal:' and 'Template:' are removed. The aim of this process is to reduce the potential noise such files would create.

Text contained within the following tags is extracted for indexing:

- Title: tag-element with name-attribute 'title' within the metadata-template (TITLE)
- Heading 1: heading-element with level-attribute '2' (H2)
- Heading 2: heading-element with level-attribute '3' (H3)
- Category: property-element with name-attribute 'type' (CAT)

Common headings such as 'References', 'External links' and 'See also' are excluded from the index.

An example of a document ready for indexing by Indri:

```
<DOC>
<DOCNO>1x6bx0ax212420</DOCNO>
<TEXT><TITLE>The Scream</TITLE><H2>Sources of inspiration</H2>
<H2>Thefts</H2><H2>In popular culture</H2><H2>Gallery</H2>
<H3>Depersonalization disorder</H3><CAT>1893 paintings</CAT>
<CAT>Edvard Munch paintings</CAT><CAT>Expressionist paintings</CAT>
<CAT>Modern paintings</CAT><CAT>Symbolist paintings</CAT></TEXT>
</DOC>
```

Our aim is to use the link structure within the Wikipedia articles to enhance the initial search. It was therefore necessary to build a hash that links the title of the article to the file location. A Perl hash with this structure is used:

```
$fileHash{'La_Concorde'} = '1x6bx0ax265017';
$fileHash{'Embassy_of_Barbados_in_Washington,_D.C.'} = '1x6bx0ax31828384';
$fileHash{'Bosnia_and_Herzegovina_Hockey_League'} = '1x6bx0ax25617492';
```

```
$fileHash{'The_Scream'} = '1x6bx0ax212420';
$fileHash{'Belgium_at_the_1924_Summer_Olympics'} = '1x6bx0ax7521518';
$fileHash{'Lambrini'} = '1x6bx0ax5264971';
```

The query against the large index is limited to the text of the description-element of the topic.

An example of a search:

```
<text>
#weight(
10.0 #combine(learn.title major.title bicycle.title
        races.title multi.title affairs.title tour.title
        de.title france.title runs.title milan.title
        san.title remo.title )
 5.0 #combine(learn.cat major.cat bicycle.cat races.cat
        multi.cat affairs.cat tour.cat de.cat france.cat
        runs.cat milan.cat san.cat remo.cat )
 2.5 #combine(learn.h2 major.h2 bicycle.h2 races.h2
                multi.h2 affairs.h2 tour.h2 de.h2 france.h2
                runs.h2 milan.h2 san.h2 remo.h2 )
 1.0 #combine(learn.h3 major.h3 bicycle.h3 races.h3 multi.h3
                affairs.h3 tour.h3 de.h3 france.h3 runs.h3
                milan.h3 san.h3 remo.h3 ) )
 </text>
```

The fields title, cat, h2 and h3 where weighted in falling importance. The top 1,000 results are returned.

Stage two In-links to the articles had the following mark-up:

```
<property name='http://dbpedia.org/ontology/wikiPageWikiLink'>
    <subject name='http://dbpedia.org/resource/List_of_paintings_by_Edvard_Munch'>
    </subject>
</property>
<property name='http://dbpedia.org/ontology/wikiPageWikiLink'>
    <subject name='http://dbpedia.org/resource/Culture_of_Norway'></subject>
</property>
<property name='http://dbpedia.org/ontology/wikiPageWikiLink'>
    <subject name='http://dbpedia.org/resource/Silence_%28Doctor_Who%29'></subject>
</property>
```

While out-links:

```
<property name='http://dbpedia.org/ontology/wikiPageWikiLink'>
    <object name='http://dbpedia.org/resource/Pastel'></object>
</property>
```

```
<property name='http://dbpedia.org/ontology/wikiPageWikiLink'>
    <object name='http://dbpedia.org/resource/Tempera'></object>
</property>
<property name='http://dbpedia.org/ontology/wikiPageWikiLink'>
    <object name='http://dbpedia.org/resource/Oil_painting'></object>
</property>
```

All the in and out links from the 1,000 articles in the result for a topic are stored in a single array together with their frequencies. The 500 most popular links are added to the initial result set.

The combination of the original articles and the most popular linked articles are then indexed. This new index is based on the entire text of the article, i.e. the tags are removed. An example of a document ready for indexing:

```
<DOC>
<DOCNO>1x6bx0ax212420</DOCNO>
<TEXT>
 212420 The Scream The Scream disambiguation The Scream
     jpg 220px The Scream Norwegian Skrik Edvard Munch
     1893 Oil painting Oil tempera and pastel on cardboard
91 73 5 Oslo National Gallery of Norway National Gallery
    The Scream Norwegian Skrik created in 18931910 The
    Scream returns damaged but younger News com au 2008
05 21 is the title of expressionism expressionist
    painting s and prints in a series by Norway Norwegian
    artist Edvard Munch showing an agonized figure against
     a  blood red sky The landscape ...
 </TEXT>
 </DOC>
```

The new smaller index, approximately 1,400 articles, was searched to obtain the final ranking.

4.3 Alternative strategies

The chosen strategy is only one of many. Future experiments could study the importance of these components in the retrieval algorithm:

- Alternative weighting of the indexed fields in the initial (stage one) search
- More fields or different fields in the initial index
- Finding related articles using only in links or out links
- A smaller or larger indexing of the articles in stage two. The index could be enhanced with Google data on incoming link texts. Freebase and other linked data sources could also be used to enhance the index.
- Only the description was used to represent the query. Would additional fields from the topic give improved results?

5 Discussion, Limitation and Further Research

At the same time that the book world becomes more and more digital,as old books are being digitized and new books are increasingly published digitally, information not published in book format becomes more and more "semantic" in the sense that data pieces (as opposed to exclusively documents in the web's first years) are linked together and made available. These two parallel development entail great opportunities in the exploitation of book material for different purposes, of which the topic of this paper is one example.

This paper provides an example of the possibilities and the challenges. Whereas "WordNet specificity", here representing content independent linguistic semantic, is one simple example of information that can be used to systematically extract semantics from written content, other much larger and much more complicated sources of semantics, the semantic web and linked data, are waiting to be used in a similar (or related) way. To explore these possibilities we will need to experiment with more modern texts than what our present test collection contains.

To judge by the results of the runs presented here, this path of research, though promising, still requires a lot of modification and calibration.

Exploring the semantics of a page in a basically statistical manner may be seen as a superposition of independent components. Counting occurrences of special words is one component on which we superimpose the detection of noun specificity. The treatment using WordNet represents further progress from the 2010 experiments, but is still rudimentary. Nouns are currently the only word-class we are treating, using only level of specificity. trying to detect classes nouns using the lateral structure of synsets may be another path to follow. It is also conceivable that treating of other word classes, primarily verbs, might contribute to the treatment. Verbs are more complicated than nouns in WordNet and such treatment will be more demanding.

Utilizing digital books poses new challenges on information retrieval. The mere size of the book text poses both storage, performance and content related challenges as compared to texts of more moderate size. But the challenges are even greater if books are to be exploited not only for finding facts, but also to support exploitation of knowledge, identifying and analyzing ideas, a.s.o.

This article represents work in progress. We explore techniques gradually in an increasing degree of complexity, trying to adapt and calibrate them.

Even though such activities may be developed and refined using techniques from e.g. Question Answering[4], we suspect that employing semantics-aware retrieval [5,6], which is closely connected to the development of the Semantic Web [7] would be a more viable (and powerful) path to follow.

One obstacle particular to this research is the test collection. Modern ontologies code facts that are closely connected to the modern world. For example the Yago2 [8] ontology, that codes general facts automatically extracted from Wikipedia, may be complicated to apply to an out-of-copyright book collection

emerging from academic specialized environments. But this is certainly a path to follow.

6 Conclusion

This article is a further step in a discussion about semantics-aware retrieval in the context of the INEX book track. Proving (or confirmation or support) of factual statements is discussed in light of some rudimental retrieval experiments incorporating semantics. We also discuss the task of proving statement, raising the question whether it is classifiable as a semantics-aware retrieval task. Results are highly inconclusive.

References

1. Kazai, G., Koolen, M., Kamps, J., Doucet, A., Landoni, M.: Overview of the inex 2010 book track: Scaling up the evaluation using crowdsourcing. In: Comparative Evaluation of Focused Retrieval. Volume 6932 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2011) 98–117
2. Fellbaum, C.: WordNet : an electronic lexical database. MIT Press, Cambridge, Mass (1998)
3. Cartright, M.A., Feild, H., Allan, J.: Evidence finding using a collection of books. In: BooksOnline '11 Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing, Amherst, MA (2011) 11–18
4. VOORHEES, E.M.: The trec question answering track. Natural Language Engineering **7** (2001) 361–378
5. Finin, T., Mayfield, J., Joshi, A., Cost, R.S., Fink, C.: Information retrieval and the semantic web. In: Proc. 38th Int. Conf. on System Sciences, Digital Documents Track (The Semantic Web: The Goal of Web Intelligence). (2005)
6. Mayfield, J., Finin, T.: Information retrieval on the semantic web: Integrating inference and retrieval. In: SIGIR Workshop on the Semantic Web, Toronto. (2003)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)
8. Hoffart, J., Suchanek, F., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Special issue of the Artificial Intelligence Journal (2012)