OAUC's participation in the 2013 INEX Book and Linked-Data Tracks

Michael Preminger[1] and David Massey[1]

Oslo and Akershus University College of Applied Sciences

Abstract. In this article we describe the Oslo University College's participation in the INEX 2013 endeavor. This year we participate in the Book Track's "Prove it" (as in 2012) and Social search tasks, as well as the Linked Data track's Ad-hoc task.
In 2011 and 2012, 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, as well as some rudimentary experiments based on named entity recognition . 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, at this point, to examine the potential utility of Amazon assigned browse nodes, a hierarchy of tags assigned to item for sale purposes.
For the Linked data task, we wish to explore the possibility of using categories to detect relevant articles.

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 have also made the claim that "Prove It!" sorts within the (admittedly 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 questions, being rhetorical in part, may be augmented by the question:

- 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

Finally

- can we build a mathematical or statistical model that captures the proving properties of a document in relation to a statement?

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

1.1 Task Definition and User Scenario

The prove-it task is currently being subjected to changes, one of which is the exclusion of pages from inappropriate books from the relevant sets. This change need some rounds of crowdsourced relevance assessments, which have not take place at the time of writing. For the present paper therefore, results are based on older relevance assessment.

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 2010 we were investigating the extent to which the distribution of certain words, inflections of "to be" and "to have" in various tenses, could be an indicator of the page being a proof / confirmation of a statement.

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

During the autumn of 2012 a new round of crowdsourced assessments was applied to the collection, using pooled results from all the participants, presenting AMT-workers[1] with pages to assess.

In the present experiments, we proceed along the same lines, this time involving the use of named entity recognition.

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 matching them to named entities in the page.

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[2]. 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 proveIt! 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

[1] The crowdsourcing uses Amazon Mechanical Turk, and "workers" is the term used for people performing Human Intelligence Tasks (HIT)

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

emphases. All this advocates for use of other strategies of presenting proving results than ranked lists, but we need to conform to the standard evaluation procedure which demands ranked lists.

1.5 Indexing and Retrieval Strategies

The point of departure of the strategies discussed have been 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 that way, the "Prove It" task was different than e.g. the one referred to in [3] This year, this condition is modified, as the relevance judgements also incorporate the containing books' appropriateness for substantiating a statement.

This means that for each encountered page, the system needs to know (or be able to find) the appropriateness of the book. The simplest strategy is that the index holds (for each page) some metadata about the book. A more elaborate strategy uses the existing metadata as basic information to use when resorting to external sources for finding appropriateness information.

Indexing In line with the above, indexing should facilitate the following main aspects at retrieval time: (a) identifying relevant pages, (b)excluding those that come from a book inappropriate for the proving the statement (c) identifying the extent of usage of named entities in a page, and (d) identifying the named entities themselves. The first aspect, (a) 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, for (b), each page is also tagged by the subject headings of the book extracted from its bibliographic record. (c) is catered for by detecting and counting named entity occurrences in each page, and tagging each page by the extent of its use of named entities, and (d) is catered for by tagging and exposing each named entity in a field tag.

The index also includes tags for confirming words and specificity ([4]), to allow us to combine approaches.

Named entity discovery Named entity discovery is a natural language processing (NLP) activity. There are several publicly available NED tools. As in 2012, we are still using the opennlp package of the Apache project. The package was used with default settings (no special training to adapt to our collection), with the assumption that the big diversity of the book collection is not apt to any significant improvement with respect to the default settings.

1.6 Runs and Results

For our group it is still the case that we see the task of refuting as distinct from (and much more complicated than) the task of confirming, and we concentrate on confirming. This means that in addition to using the official qrels (where confirmation and refutal are merged) we will also use "pure confirming" qrels, from which refutal assessments are separated and filtered away.

Identifying a baseline Even though our documents are all book-pages they may vary in length, and other language attributes meaning that identifying the smoothing that gives the best baseline for the documents as a whole requires some trial and error. As in all our experiments reported in this paper, the best baseline would be identified by running a basic edition of our query types against our index using a scan of λ (JM constants: 0.1 : 0.1 : 0.9). The optimal baseline parameters are then used in the semantic variation we do on baseline retrieval when testing its effect.

Named entities: occurrence rate of named entities in the page As already indicated, rudimentary experimentation has been done with named entity recognition in this year's experimentation. The main parameter is the number of occurrence of named entities in the page. This is a very simple measure, insensitive to the length of the page itself. But as an initial gauge of the merit of name entities in detecting confirmation, we feel it is sufficient, particularly when the property will be used in combination with other properties. We simply count the number of times a location or a personal name occurs in the page, and have nine different tags to stand for 1 - 8 occurrences, and nine or more occurrences, respectively. The various runs give extra weight to pages within different range of occurrence rates.

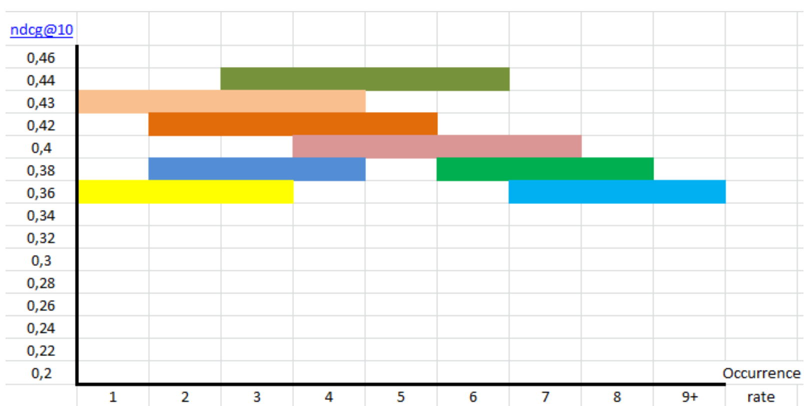

Fig. 1. Weighed up ranges of occurrence rates and their ndcg@10 score.

Referring to runs weighting named entity pages in isolation, it is difficult to pin-point a level of occurrence of named entity that stands out in performance. Named entity presence in a page does not seem to be a very good indicator of proving when used as the sole indicator.

Named entities: specific named entities occurring in the topic queries A more topic-specific approach than the former is searching the documents for occurrences of the same named entity identified in the query. Here it is not enough to match the named entity as it appears in the query, since a matching entity in a document may appear in a different form. This calls for expanding the query with as many as possible of the known forms of the entities found in it, with the hope that one of them will match the entity appearing in the page. This expansion, needless to say, is done at retrieval time. In the current experiments, we are using the MediaWiki API (ref) query and backlink facilities (ref). This is done in two steps: (a)querying Wikipedia for each of the named entities identified in the topic and identifying its preferred form and (b) applying backlinks to the preferred form to find all the forms that redirect to (a). All those forms are then used (as synonym groups for each named entity) to expand the query, placing a higher weight on retrieving documents with named entities. This means that documents (pages) having matching named entities are assumed to have a higher confirming probability with respect to the statement in question.

Appropriateness of the containing book to confirmation of the statement New to 2013, a path of research we wish to pursue is the appropriateness of the containing book for confirming a claim, so that pages from inappropriate books are not counted. As our books have MARC-records associated with them, it is natural, to this end, to extract subject heading information from the MARC records. 21500 of the book-records contain the 650 (subject added entry) tag. About 28000 records have a 65X tag (including the former), and about 35000 (including both former groups include any 6XX tag. 082 (Dewey Decimal Classification) is rarely used in these records, and is therefore of a lesser value to this task). We see particularly the 650 tags as potentially helpful for identifying the genre or main topic of a book, and including it in an index, so that each page has the subject of its containing book in an appropriate tag. Using these tags is still a problem. Books that do not have this tag, need a less obvious treatment so that their appropriateness can be assessed.

There still remains the challenge to automatically deduce from the statement which subjects are appropriate to confirm it.

1.7 Result summary and discussion

The results above indicate that there is more potential in treating occurrences of named entities occurring in the query rather than the general occurrence rate of named entities in documents. Still, more experiments are necessary.

The experimental research into the prove It! task requires, in our opinion, gradual advance in several fronts. Diverting from traditional IR, two main activities need to be refined in a semantic direction: the semantic aware search algorithms and the intellectual "relevance assessment", here understood as refining the preparation of the crowdsourcing activities. The goal must be to arrive at models that incorporate the different aspects of proving/confirming/substantiating discussed above that may predict the proving power of a document with respect to a statement.

run	ndcg@10	comment
baseline	0,6931	JM Lambda = 0.7
Named entity occurrence in text	0.4360	lambda=0.7 pages with three to six ne occ. weighted dubble
Named entity in query matched in text	0,6069	JM lambda = 0.4 Pages matching synomyms in the list weighted double

Fig. 2. preliminary results for the modes of named entity usage.

At this stage we are assessing the potential of each of the aspects of named entity discovery, to explore how each of them (and a combination of them) can contribute to a linear model that predicts the confirming power of a document.

A baseline as ours (created with smoothed language modeling algorithms taken from traditional IR) is difficult to beat, but it is nevertheless interesting to measure how closely we approach the results by above mentioned combinations.

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.

[5] has attempted to compare the utility of social data to that of DEWEY classification data (which are also available in the Amazon records). In 2012 we tried to build upon this, and looked at subject headings extracted from the MARC data. The results were not promising, and this path of research was abandoned. There seems to be a discrepancy between book search in a social context and the same search in a "content context" which the MARC-data are meant to support.

As the case is for the prove It! research discussed above, also here we need to advance gradually. Traditional IR is content-aware. Social search does not always originate in a knowledge (read:content) anomaly ([6]). Social search is necessarily related to the social context in which the search originates, we hypothesize that it is related (among other factors) to the type of books. Amazon BrowseNodes are used to organize items for sale (as opposed to the more content oriented subject headings). We preferred to use them this year, rather than the libraryThing user provided tags although the latter are definitely interesting in this context. The BrowseNodes are related to content and genre, but at the same time also to consumption, as they are created and deleted according to item demand. This means that we in general do not have a constant relation between a book and a browseNode. As a first attempt we only use a snapshot of the browseNode organization as a different form of "Subject Heading" that may (or may not) be more appropriate in nature to social book search than the latter, more LIS-oriented type.

2.2 Indexing and retrieval strategies

The collection has been loaded into a database where all types of data about each book are associated with the book's ISBN. We create an Indri index that contains two representations of each book as two documents, one containing the browseNodes and the other containing the rest of the tags under the <book> hierarchy - all the tagged XML information that is extracted from both amazon and the LT fora. The result list will be a mixture of documents of both representations, and they are collated so each original document takes the rank of the best of its representations in the list.


```
<DOC>
<DOCNO>0836921283_K</DOCNO>
<TEXT>
<book><isbn>0836921283</isbn><title>The joyful gardener (
    Essay index reprint series)</title><ean
    >9780836921281</ean><binding>Unknown Binding</binding
    ><label>Books for Libraries Press</label><listprice/><
    manufacturer>Books for Libraries Press</manufacturer><
    publisher>Books for Libraries Press</publisher><
    readinglevel/><releasedate/><publicationdate>1971</
    publicationdate><studio>Books for Libraries Press</
    studio><edition/><dewey/><numberofpages>274</
    numberofpages><dimensions><height/><width/><length/><
    weight/></dimensions><reviews/><editorialreviews/><
    images/><creators><creator><name>Agnes Edwards Rothery
    </name><role>Author</role></creator></creators><
    blurbers/><dedications/><epigraphs/><firstwords/><
    lastwords/><quotations/><series/><awards/><characters
    /><places/><subjects/><tags><tag count="1">essays</tag
    ><tag count="1">box 18</tag><tag count="1">plants</tag
    ><tag count="1">gardening</tag></tags><similarproducts
    /></book>
</TEXT>
</DOC>
<DOC>
```

```
<DOCNO>0836921283_N</DOCNO>
<TEXT>
<browseNodes><browseNode id="48">Home &amp; Garden</
    browseNode><browseNode id="1000">Subjects</browseNode
    ><browseNode id="5241">Gardening &amp; Horticulture</
    browseNode><browseNode id="5276">General</browseNode><
    browseNode id="283155">Books</browseNode><browseNode
    id="713339011">General AAS</browseNode></browseNodes>
</TEXT>
</DOC>
```

At retrieval time, doubling the weight of browseNodes can be done

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

After retrieval, the ranked lists are collated as the original book document takes the place of its higher ranked representative. In this way we can more easily isolate the contribution of the different parts of the document to the retrieval performance.

2.3 Preliminary runs and results

Preliminary runs were performed in accordance with the description in Section 2.2. Figure 3 summarizes the results. As for now, the queries are quite "naive" as related to the Browsenodes, and no expansion is taking place. a natural path to consider is expanding the browsenodes with synonyms before indexing them, so that the queries are not dependent on having the exact wording of the browsenode for matching a document containing it. Some work needs also to be done to find the exact average net relative weight of the document representations in order to

3 The linked data track, the Ad-Hoc task

3.1 Introduction

The purpose of the linked data track is to find out how techniques from semantic web / linked data can be used to improve and enhance retrieval of Wikipedia

run	ndcg@10	comment
1*browseNodes 1*(rest of book)	0,0172	Default smoothing
1.1*browseNodes 1*(rest of book)	0,0155	Default smoothing
1*browseNodes 1*(rest of book)	0.0408	jm, lambda=0.4
1.1*browseNodes 1*(rest of book)	0,0354	jm, lambda=0.4

Fig. 3. preliminary results for the Social search book track

articles. The data collection is an XML'ified version of a Wikipedia subset (about 4.1M articles), where RDF-properties from DBPedia and YAGO, are combined with the article text.

The idea behind 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 a small selection of the file's contents. For each topic, categories found in the articles from the initial search result (1000 articles) are extracted and matched against the topic. If the similarity is above a given threshold the articles related to the category are inserted at the top of the final result list.

Wikipedia categories have been used to enhance ad hoc retrieval [7] and as the basis for entity ranking [8]. Our focus on categories is based on the idea that broad queries will be satisfied by categories while more specific, entity level queries will be satisfied by the stage one index.

3.2 Indexing and retrieval strategies

Stage one In an effort to reduce noise, files that are not considered articles, such as metadata on images and deleted articles are ignored. The original Wikipedia articles are reduced to skeleton documents consisting of a selection of the semantic markup. In our 2012 run only title, headings and categories where indexing. In 2013 the extraction process has been expanded to include the following six fields:

- the title (TITLE)
- the infobox type (INFO-BOX)
- the title of other entities that are linked to (out links) from the article text (ENTITY)
- the Dublin Core subjects from the DBPedia properties (DC-SUBJECT)
- the RDF types from the DBPedia properties (RDF-TYPE)
- the RDFS labels from the Yago properties (ALT)

As an example, the Wikipedia article on Edvard Munch, ready for indexing by Indri (please note that the example is abridged due to space limitations):

```
<DOC>
<DOCNO>1x1axf2x9779</DOCNO>
<TEXT><TITLE>Edvard Munch</TITLE>
<ENTITY>mona lisa</ENTITY>
<ENTITY>degenerate art</ENTITY>
<ENTITY>auguste rodin</ENTITY>
<DC-SUBJECT>expressionist painters</DC-SUBJECT>
<RDF-TYPE>symbolist painters</RDF-TYPE>
<DC-SUBJECT>norwegian artists</DC-SUBJECT>
<ALT>edvvard munch</ALT>
<ALT>edward munch</ALT>
<ALT>edvard</ALT></TEXT>
</DOC>
```

The ALT field combines alternative forms of the article title (re-directs), e.g. "edvvard munch" and "edward munch", and disambiguations, e.g. "edvard". DBPedia and Yago properties are also generated from the same structured parts of the Wikipedia article [9] and this can lead to duplicate values. A de-duplication routine is therefore utilized prior to indexing. Krovetz stemming and the default 418 word stop word list are used under indexing.

An unweighted search for a query generated from the unique words combined from the topic title and description using two-stage smoothing with a lambda value of 0.9 creates the initial ranked list which forms the baseline for stage two.

Stage two Each document from the result list is fetched and the subjects (Wikipedia categories) are extracted. The category name is matched against the topic title. Prior to the matching the topic title is expanded with synonyms from Wordnet and a small list of stopwords (in, on, of, the, and, by, to) are removed. The Levershtein edit distance measure is used to find the lexical similarity between topic and category. The best category word is found for each synonym group and the total is normalised by the category length. Only categories with a similarity score above a 0.75 threshold result in the insertion of associated articles. Articles from the original ranked list, that are not assigned to the selected category are pushed down the list.

Only six of the 144 topics resulted in categories that are above the chosen threshold (compared to six of the 30 evaluated topics in 2012) which suggests that the 2013 topics don't have the same broad nature of the previous year.

We are awaiting results for our 2013 run.

3.3 Alternative strategies

Future work is required to examine why only such a small number of the topics found an appropriate category. The quality of category assignment in Wikipedia is a further issue that could be investigated and the possibility of assigning categories automatically. Exploiting the hierarchical nature of the Wikipedia category structure and finding links to additional subject descriptions are of interest.

4 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"([4], here representing content independent linguistic semantic, and named entity recognition are simple examples 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 or treatments. Particularly in the ProveIt! task, we wish to pursue the treatments from [10] and [4], along with this year's further. As for the Social search, the combination of short queries and short documents seem to be very sensitive to smoothing parameters. More work is needed to find out if the direction is worth pursuing.

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[11], we suspect that employing semantics-aware retrieval [12,13], which is closely connected to the development of the Semantic Web [14] 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 [15] 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.

5 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. Preminger, M., Nordlie, R.: Ouc's participation in the 2011 inex book track. In: Comparative Evaluation of Focused Retrieval. Lecture Notes in Computer Science, Springer Berlin / Heidelberg
5. ogers, T., C.K.W..L.B.: Rslis at inex 2011: Social book search track. In: Comparative Evaluation of Focused Retrieval. Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2010) 49–59
6. Belkin, N.J., Oddy, R., Brooks, H.: Ask for information retrieval: Part i. background and theory. Journal of Documentation **38**(2) (1982) 61–71
7. Kaptein, R., Koolen, M., Kamps, J.: Using wikipedia categories for ad hoc search. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM (2009) 824–825
8. Thom, J.A., Pehcevski, J., Vercoustre, A.M.: Use of wikipedia categories in entity ranking. arXiv preprint arXiv:0711.2917 (2007)
9. Auer, S., Lehmann, J.: What have innsbruck and leipzig in common? extracting semantics from wiki content. In: The Semantic Web: Research and Applications. Springer (2007) 503–517
10. Preminger, M., Nordlie, R.: Ouc's participation in the 2010 inex book track. In: Comparative Evaluation of Focused Retrieval. Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2010) 164–170
11. VOORHEES, E.M.: The trec question answering track. Natural Language Engineering **7** (2001) 361–378
12. 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)
13. Mayfield, J., Finin, T.: Information retrieval on the semantic web: Integrating inference and retrieval. In: SIGIR Workshop on the Semantic Web, Toronto. (2003)
14. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)

15. 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)