Evaluating Web-as-corpus Topical Document Retrieval with an Index of the OpenDirectory

Clément de Groc*   Xavier Tannier†

*Syllabs
26 rue Notre Dame de Nazareth, 75003 Paris, France
cdegroc@syllabs.com

†LIMSI-CNRS
Univ. Paris-Sud, 91403 Orsay, France
xtannier@limsi.fr

Abstract

This article introduces a novel protocol and resource to evaluate Web-as-corpus topical document retrieval. To the contrary of previous work, our goal is to provide an automatic, reproducible and robust evaluation for this task. We rely on the OpenDirectory (DMOZ) as a source of topically annotated webpages and index them in a search engine. With this OpenDirectory search engine, we can then easily evaluate the impact of various parameters such as the number of seed terms, queries or documents, or the usefulness of various term selection algorithms. A first fully automatic evaluation is described and provides baseline performances for this task. The article concludes with practical information regarding the availability of the index and resource files.

Keywords: Web-as-corpus, Information Retrieval, Evaluation

1. Introduction

Specialized terminologies and corpora are key resources in applications such as machine translation or lexicon-based classification. However, they are also expensive to develop. Using the Web as a corpus (Kilgarriff and Grefenstette, 2003) helps constructing semi-automatically such resources from the Web. Search engine-based methods in particular use topic-specific queries to discover and retrieve specialized documents from the Web.

The BootCaT procedure (Baroni and Bernardini, 2004) is a widely used method to bootstrap specialized corpora and terms using topic-specific queries to a Web search engine. The procedure requires only a small set of seed terms as input. The terms are mixed into queries and submitted to a Web search engine. The Top-M documents are then fetched yielding a small corpus. More terms can then be extracted from the corpus and used to build a bigger corpus in an iterative way. Evaluation of the resulting corpora is usually done manually, for a particular language and topic (see for example (Baroni and Ueyama, 2004; Leturia et al., 2008; Baroni and Bernardini, 2004)).

As search-engine based Web-as-corpus approaches are getting more and more used, we believe that there is a need for an automatic, reproducible and robust evaluation protocol.
The evaluation should be automatic so we can observe the impact of the numerous parameters (Kilgarriff et al., 2011) on the search results: query size, number of queries, term extraction algorithm, number of webpages fetched by query, etc. The evaluation should be reproducible by relying on a static collection of pages instead of the Web that is constantly changing. Finally, as some domains have very clear and unambiguous terminologies (*e.g.* biology or cooking), while others (*e.g.* sociology) share a large part of their vocabulary with common language (Kluck and Gey, 2001), our evaluation should also cover a large panel of topics.

In the remaining of this article, we describe a novel evaluation protocol and resource based on the OpenDirectory[1]. After fetching all Web pages mentioned in the English part of the OpenDirectory, we index them in an Open Source search engine. We then extract a set of seed terms automatically for 340 topics of the second level of the OpenDirectory and provide first results regarding the impact of query size on precision and recall. Finally we conclude with practical information about the availability of the index and resource files.

2. The OpenDirectory corpus

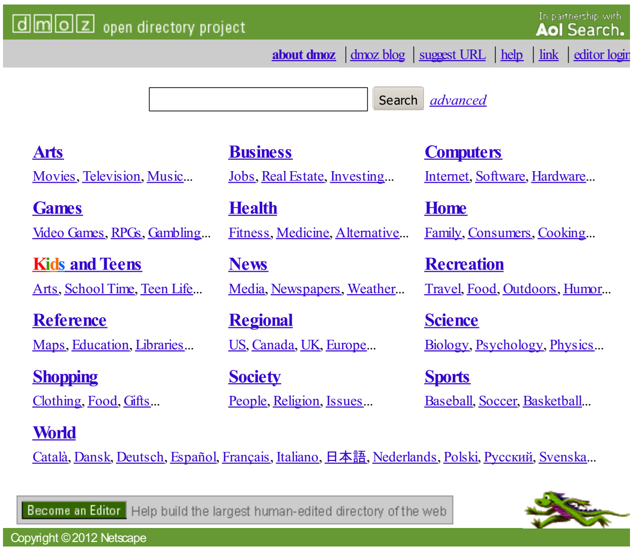
The OpenDirectory, sometimes called ODP or DMOZ (*Directory Mozilla*), is a directory of Web sites maintained by a community of volunteers. As of September 2013, the directory is composed of 5,262,071 records in 88 languages. In the ODP, Web sites are classified in a topical thesaurus (hierarchical tree of topics) and annotated with a short description of their content. We present the ODP home page showing sample topics in Figure 1. The repository is updated regularly and is available for download in the RDF format (XML). Two sample entries are shown in Figure 2.

Our repository dates back from September, 15th 2011 and does not include the `Adult`[2] and `Kids_and_Teens`[3]

[1] `http://www.dmoz.org`

[2] `http://www.dmoz.org/guidelines/adult/`

[3] `http://www.dmoz.org/guidelines/kguidelines/`


dmoz open directory project
In partnership with Aol Search.
about dmoz | dmoz blog | suggest URL | help | link | editor login
Search
advanced
Become an Editor
Help build the largest human-edited directory of the web
Copyright ©2012 Netscape

Arts Movies, Television, Music...	**Business** Jobs, Real Estate, Investing...	**Computers** Internet, Software, Hardware...
Games Video Games, RPGs, Gambling...	**Health** Fitness, Medicine, Alternative...	**Home** Family, Consumers, Cooking...
Kids and Teens Arts, School Time, Teen Life...	**News** Media, Newspapers, Weather...	**Recreation** Travel, Food, Outdoors, Humor...
Reference Maps, Education, Libraries...	**Regional** US, Canada, UK, Europe...	**Science** Biology, Psychology, Physics...
Shopping Clothing, Food, Gifts...	**Society** People, Religion, Issues...	**Sports** Baseball, Soccer, Basketball...
World Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...		

Figure 1: DMOZ home showing sample topics.

URL	`http://www.lrec-conf.org/`
Title	LREC Conferences
Topic	`Science/Social_Sciences/Linguistics/Computational_Linguistics/Conferences`
Desc.	The International Conference on Language Resources and Evaluation is organised by ELRA biennially with the support of institutions and organisations involved in HLT. LREC Conferences bring together a large number of people working and interested in HLT.

`http://www.lonelyplanet.com/`
Lonely Planet
`Recreation/Travel/Guides_and_Directories`
Offers travel advice, detailed maps, travel news, popular message boards and health information. Also lists information and updates regarding guidebooks.

Figure 2: Two sample entries of the OpenDirectory.

categories, which are distributed in a separate archive. We focus on the English part of the directory, meaning that all entries under the `World` category are discarded.

The OpenDirectory solely provides URL of Web pages but not web pages content. Therefore, we have to download the Web pages manually, on top of the RDF dump. Downloading several millions of web pages and processing them requires a certain technicality (respect robots.txt, follow redirections, handle encodings, distribute fetching...), which, once achieved, gave us 2,339,125 Web pages over a total of 2,463,769 URLs in English. The small percentage of lost pages is distributed among all categories and does not impact our conclusions.

A preliminary study of the directory showed that some of the categories of the ODP do not convey a topical nature. Therefore, we apply the following filtering process to improve the quality of the corpus. We exclude all categories under the `Regional` topic that contains a geographical (non-topical) classification of web sites[4]. We also exclude a number of categories corresponding to genres of web pages and remove `Chats_and_Forums`, `Directories`, `FAQs,_Help,_and_Tutorials`, `Magazines_and_E-zines`, `Mailing_Lists`, `News`, `News_and_Media`, `Personal_Pages`, `Search_Engines`, and `Weblogs` categories, as well as purely organizational categories (`By_Culture`, `By_Region`, `By_Type`, ...), "letters" categories (list of sportsmen starting with letter *X*) and "dates" categories (`Roland_Garros/2005,2006,2007,...`).

Finally, we limit the scope of our experiments to topics of the second level of the OpenDirectory with at least 50 documents.

3. Indexing in Lucene

We now index Web pages in Apache Lucene[5] (Bialecki et al., 2012), a very efficient, and widely used Open Source search engine. We apply a rudimentary cleaning algorithm on webpages before indexing: we parse the web pages using the TagSoup HTML parser[6] and remove style sheets, dynamic scripts, comments and HTML markup before indexing the remaining text with Lucene. We use Lucene's default Analyzer (*StandardAnalyzer*) that applies a simple tokenization, lower-case transformation, and stopwords removal. Besides indexing documents content, we also store their URL, topic(s), description, and title in the index.

4. Automatic seed terms extraction

We now focus on the automatic selection of seed terms. Indeed, there is as mush as 340 categories at the second level of the OpenDirectory, hence the need for an automatic way to select seed terms for each category.

We rely on *topic descriptions* (Srinivasan et al., 2005) to select seed terms. Topic descriptions are built by concatenating the manually entered descriptions of all web pages under a certain topic of the OpenDirectory. We then model those texts as bag-of-words and apply a TermHood (Kageura and Umino, 1996) measure to extract relevant and discriminative seed terms. We use the *tfidf* heuristic which provided more relevant keywords than log odds ratio (Everitt, 1992) or weirdness (Ahmad et al., 1999) in our experiments.

The *tfidf* variant we used is the *ntc* (Manning et al., 2008, chap. 6, p. 128) and is defined as follows:

$$\mathrm{tfidf}_{t,d} = \mathrm{tf}_{t,d} \times \log \mathrm{idf}_t \qquad \mathrm{idf}_t = \frac{D}{\mathrm{df}_t}$$

[4] `http://www.dmoz.org/docs/fr/guidelines/subcategories.html#regional`

[5] Version 4.7, `http://lucene.apache.org/`

[6] `http://home.ccil.org/~cowan/XML/tagsoup/`

where $\mathrm{tf}_{t,d}$ is the number of occurrences of term t in the topic description d, df_t the number of descriptions where term t appears and D the total number of topic descriptions in the collection.

We present the top-10 terms for five topics according to their *tfidf* weights in Table 1.

5. Evaluation

In this section, we evaluate the quality of results retrieved by querying the ODP index using the automatically extracted seed terms. The evaluation protocol is as follows:

1. For each topic, we select the top-N automatically extracted seed terms.
2. We generate all tuples from those terms with varying size, from 1 to 7.
3. We create queries from those tuples using a conjunction operator (*AND*).
4. We submit all queries to the OpenDirectory search engine and retrieve (at most) M documents for each query.
5. We merge those documents into a corpus and evaluate the Precision, Recall and F_1-measure. Note that documents fetched by each query might overlap and that we expect a low recall since we limit the number of queries and documents fetched.

In our experiment, we fixed N and M to 10. We consider a document relevant if its topic in the ODP is the same than the query's topic. As can be seen from Table 2, single term queries are too ambiguous and offer low precision and recall. Precision increases with the size of the queries, while recall is maximal for queries of size 3. A more in-depth study of the results show two main sources of error:

- Some keywords seem valid but aren't discriminative enough.
- A few topics are very ambiguous (`Arts/Crafts` and `Shopping/Crafts`, `Arts/Video` and `Arts/Movies`).

Both issues might be tackled by refining the set of topics considered, using a different TermHood measure or manually validating the sets of seed terms.

We foresee many more hypotheses that could be studied using our resource. To name a few, we plan to tackle the following questions:

- How does the number of seed terms, queries and documents interact together? Which one should we favour and in which situation?
- Which TermHood measure should we use?
- Can we bias the random creation of tuples towards more relevant queries?

Table 2: Macro averaged precision, recall and F_1-score for various tuple (query) sizes. We also present the number of queries issued and average number of documents fetched for each topic.

Size	Nb queries	Nb docs	P	R	F_1
1	10	96.7	0.263	0.055	0.065
2	45	264.3	0.356	0.155	0.149
3	120	337.8	0.367	**0.173**	**0.165**
4	210	288.5	0.382	0.144	0.151
5	252	197.4	0.399	0.099	0.120
6	210	115.2	0.419	0.061	0.085
7	120	58.0	**0.439**	0.032	0.052

6. Conclusion

In this article, we have described a novel protocol to evaluate search engine-based Web-as-corpus approaches in an automatic, reproducible and robust way. Our method is based on the indexing of an annotated subset of the Web (the OpenDirectory) in an Open Source search engine (Lucene).

We ran a first experiment to evaluate the impact of query size on precision and recall. We validated experimentally that precision increases with query size, while recall decreases for queries composed of more than 3 terms. However, even for very large queries of 7 terms, precision remains below 0.5 while recall drops drastically.

In our future work, we plan to investigate more hypotheses using our dataset such as how the number of seed terms, queries and documents interact together and which should be modified depending on the situation. Another promising lead could be to use this dataset to measure terms discriminative power and therefore select better seed terms and queries.

We invite researchers to contact both authors to gain access to the Lucene Index, OpenDirectory files (RDF dump and Web pages corpus) or Java code.

7. References

K. Ahmad, L. Gillam, and L. Tostevin. 1999. University of surrey participation in trec 8: Weirdness indexing for logical document extrapolation and retrieval (wilter). In *The Eighth Text REtrieval Conference (TREC-8)*.

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the LREC 2004 conference*, pages 1313–1316.

M. Baroni and M. Ueyama. 2004. Retrieving japanese specialized terms and corpora from the world wide web. In *Proceedings of KONVENS*, pages 13–16.

A. Bialecki, R. Muir, and G. Ingersoll. 2012. Apache lucene 4. In *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval.*

B.S. Everitt. 1992. *The analysis of contingency tables*, volume 45. CRC Press.

Table 1: Top 10 terms with highest *tfidf* for five topics at the second level of the ODP.

Business/Energy	Computer/A.I.	Home/Cooking	Science/Math	Society/Paranormal
solar	neural	recipes	mathematics	psychic
energy	reasoning	recipe	mathematical	readings
gas	algorithms	servings	algebraic	paranormal
oil	bayesian	broth	algebra	clairvoyant
biodiesel	ai	sauce	theory	tarot
electric	networks	cheese	geometry	ufo
electricity	computational	chicken	math	ghost
drilling	intelligence	cream	equations	psychics
water	learning	recipesource	calculus	intuitive
wind	machine	onion	department	haunted

K. Kageura and B. Umino. 1996. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

A. Kilgarriff, PVS Avinesh, and J. Pomikálek. 2011. Comparable corpora bootcat. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex*, pages 122–128.

M. Kluck and F. Gey. 2001. The domain-specific task of clef - specific evaluation strategies in cross-language information retrieval. *Cross-Language Information Retrieval and Evaluation*, pages 48–56.

I. Leturia, I. San Vicente, X. Saralegi, and ML de Lacalle. 2008. Collecting basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Web as Corpus 4 workshop Proceedings*, pages 40–46.

C.D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to information retrieval.* Cambridge University Press Cambridge.

P. Srinivasan, F. Menczer, and G. Pant. 2005. A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3):417–447.