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

Clément de Groc
Syllabs & Univ. Paris-Sud, Orsay, France
cdegroc@limsi.fr

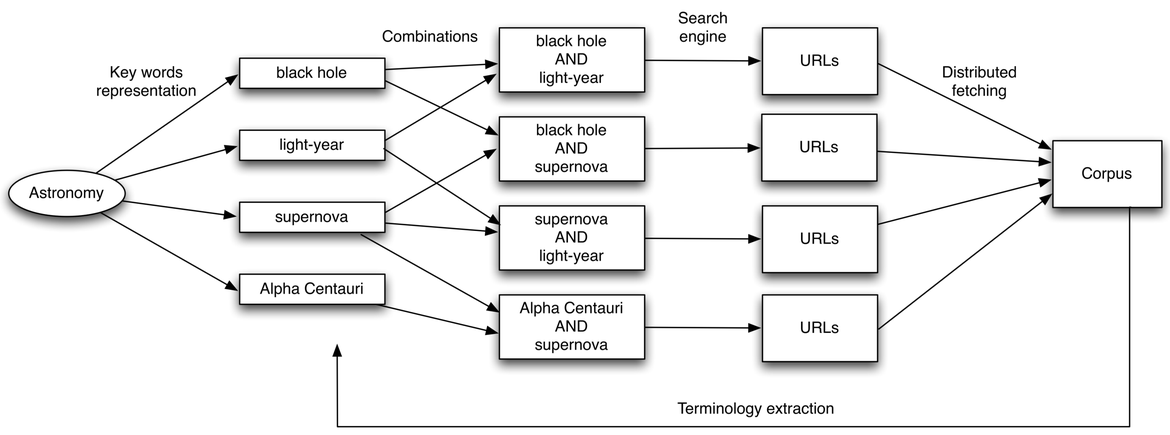
Xavier Tannier
LIMSI-CNRS, Univ. Paris-Sud, Orsay, France
xtannier@limsi.fr

Motivation

1. Topical Document Retrieval

Goal:
- Create domain-specific search engines
- Create specialized terminologies and corpora from the Web

Selected approach:
- Piggyback on commercial Web search engines (BootCaT [1])



2. Evaluation

Current methods:
- Evaluate manually documents and terms
- Usually applied to a single topic
- Fix parameters empirically after a few tests (tuple length, no. tuples, no. documents per query)

Cons:
- Non reproducible: the Web and search engines are changing
- Costly: can't easily evaluate the impact of various parameters
- Domain-specific: can't conclude about results on a different domain

Proposal:
- Reproducibility: Rely on a fixed OpenDirectory corpus
- Fast: Index locally in an open source search engine
- Robust: Evaluate on a wide panel of topics

Evaluation with an index of the OpenDirectory [3]

1. Indexing the OpenDirectory corpus

- OpenDirectory: Large topically organized thesaurus of Web sites.
- Remove non-topical categories (Regional, genres) and fetch Web pages.
- Remove HTML markup and index in Apache Lucene [4].
- We only consider English language and second level topics in our work. Method remains the same for finer-grained topics and other languages.

Our index contains more than 900k Web pages classified in 340 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 […]
	`http://www.aclweb.org/index.php`
	The Association for Computational Linguistics: Conferences
	Science/Social_Sciences/Linguistics/Computational_Linguistics/Conferences
	Information on upcoming ACL (and associated) conferences. Also archives mirrors of past conferences.

Figure 1: *Two sample entries from the OpenDirectory*

2. Evaluating with the OpenDirectory index

Method:
- Automatically find seed terms for each of the 340 topics.
- Combine as queries and evaluate proportion of relevant documents retrieved.

Automatic seed terms extraction:
- Create *topic descriptions* [2]: concatenation of all site descriptions (*fig.1*) in a topic.
- Tokenize and order by *tf.idf* statistic on all 340 topic descriptions (*fig. 2*).
  - *tf*: term frequency in one topic description.
  - *idf*: inverse document frequency over all topic descriptions.

Business/Energy	Computer/A.I.	Science/Math
solar	neural	mathematics
energy	reasoning	mathematical
gas	algorithms	algebraic
oil	bayesian	algebra
biodiesel	ai	theory
electric	networks	geometry
electricity	computational	math
drilling	intelligence	equations
water	learning	calculus
wind	machine	department

Figure 2: *Top-10 seeds terms extracted automatically for 3 topics.*

3. First evaluation

For a topic:
- Select Top-10 seed terms.
- Generate tuples with varying size, from 1 to 5.
- Create queries from tuples using conjunction (AND) operator.
- Submit queries to Lucene and collect Top-10 documents.

Evaluate Precision, Recall and F1-score by looking up documents topic.
Evaluate macro-averaged measures over all topics *(fig. 3)*.

Size	No. queries	No. docs	Precision	Recall	F1-score
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

Figure 3: *Macro-averaged precision, recall and f1-measure of topical document retrieval.*

Contact us to get access to the Lucene index of the OpenDirectory!

We provide the DMOZ corpus, seed terms and source code as well.

References

[1] M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In Proceedings of the LREC 2004 conference, pages 1313–1316.
[2] P. Srinivasan, F. Menczer, and G. Pant. 2005. A general evaluation framework for topical crawlers. Information Retrieval, 8(3):417–447.
[3] `http://www.dmoz.org`
[4] `http://lucene.apache.org`