Visualizing the Problems with the INEX Topics

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Maria del Rocio Gomez Crisostomo
Universidad de Extremadura
Badajoz, Spain
mrgomcri@alcazaba.unex.es

Mounia Lalmas
Department of Computer Science
University of Glasgow
Glasgow, Scotland
mounia@dcs.gla.ac.uk

ABSTRACT
Topics form a crucial component of a test collection. We show, through visualization, that the INEX 2008 topics have shortcomings, which questions their validity for evaluating XML retrieval effectiveness.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models, Search process*.

General Terms
Reliability, Experimentation, Human Factors, Verification.

Keywords
INEX, XML-IR, element retrieval. IR methodology.

1. INTRODUCTION
Hawking *et al*. [2] outline 12 desirable features of search engine evaluation of which 4 apply to topics, paraphrasing: (1) Many topics should be used; (2) When comparing maximal (not typical) effectiveness, the full range of search facilities should be exploited; (3) Topics should be representative of genuine user needs; (4) Topics should represent the full range of information needs. Although originally formulated for web-IR, these features apply equally well to any form of text-IR evaluation.

We ask if the INEX 2008 topics satisfy these features, and one other – (5) Topics should be independent from each other. It is shown visually and through statistical analysis that the INEX 2008 topics only satisfy FEATURE 1. This provides further evidence of the Trotman & Lalmas [6] claim that users, in our case INEX participants, are particularly bad at specifying structural hints, and the first evidence to refute the Lehtonen [4] claim that this is a consequence of the (IEEE) document collection used at INEX up to 2005.

2. MANY TOPICS (THE COLLECTION)
We use the INEX Wikipedia collection of 659,388 documents, the INEX 2008 topics (v4) consisting of 135 participant submitted topics (the INEX set) and 150 topics drawn from a New Zealand high school proxy log (the Proxy set). The INEX topics contain two fields of interest: the CO query and the CAS query; the former are typical of queries seen by online search engines, the latter additionally contain support and target structural hints. CO queries are used throughout, except in section 3 which uses CAS.

It is typical for 50 topics to be used at TREC and for about 100 at INEX. Buckley & Voorhees [1] suggest that 50 topics is sufficient for a low error rate (< 3%) in most measures. It is reasonable to conclude that FEATURE 1 is satisfied.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

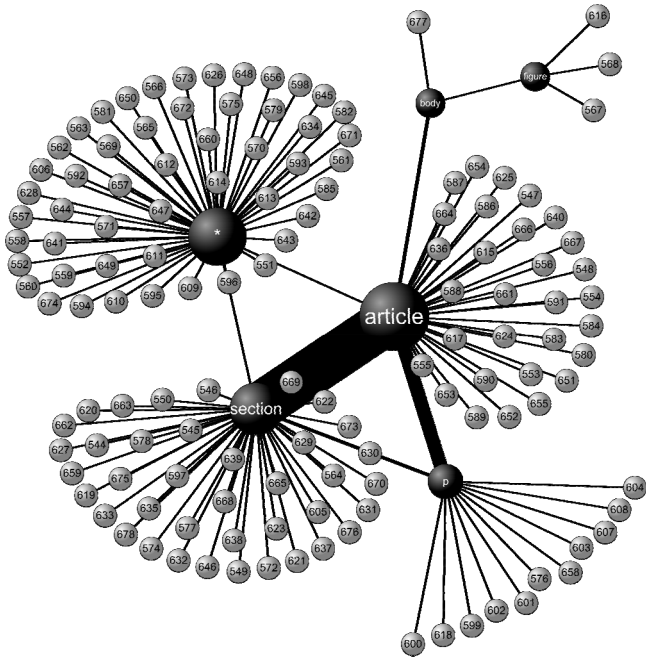

Figure 1: The structural target elements (black) in the topics (grey) show essentially no use of semantic structural hinting

3. FULL RANGE OF FACILITIES
The facilities provided by the INEX CAS query language are: phrase, restriction (+,-), target & support structural hints, Boolean, and arithmetic filters. Table 1 shows the frequency of use of these features. Superficially, most facilities are used.

Table 1: Use of each search facility in the INEX 2008 topics

Feature	#Q	%	Feature	#Q	%
Phrase	13	10%	+/-	18	13%
Support	7	5%	Target	86	64%
Boolean	25	19%	Arithmetic	0	0%

Figure 1 shows which topics (gray) target which elements (black). Black sphere size is a function of the observed frequency in the topics whereas line width is a function of relationship frequency. Only 5 (of a possible 1,241) tags are used (0.4%): `article`, `body`, `section`, `p`, and `figure` (and unspecified, `*`). Only 3 of the topics exercise element-retrieval (targeting `figure`). Targeting `article` or `body` is document-retrieval; targeting `p` is passage retrieval; targeting `section` is specifying a result

size. Lehtonen observed structural hints in the IEEE collection targeting result size; this is seen in the Wikipedia topics too. Although the topics test many facilities, the use of structural hints for XML-retrieval is not well represented in the topic set, so it is reasonable to conclude that FEATURE 2 is not satisfied.

4. GENUINE USER NEEDS

Figure 2 shows the Zipfian distribution of terms in the Wikipedia document collection and a sliding window count of the number of topic terms in each set. Search terms in both sets tend to occur in tens of thousands of documents. Topic lengths are plotted in Figure 3. Length is inversely proportional to topic frequency in the Proxy set, but in the INEX set there is a preference for topics of length 3. FEATURE 3 is not well satisfied.

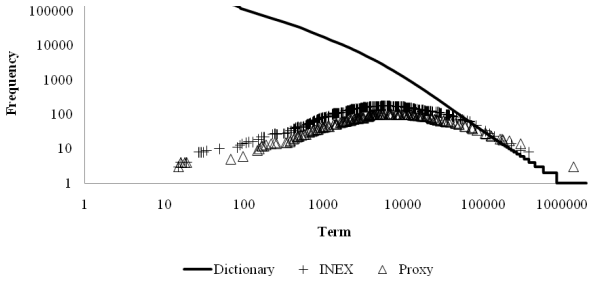

Figure 2: Term frequencies in the two topic sets are similar

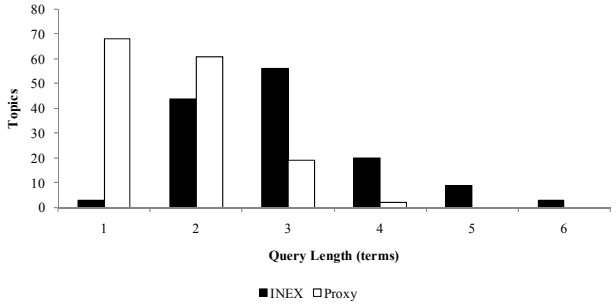

Figure 3: Query lengths in the two topic sets differ

5. TOPIC INDEPENDENCE

Figure 4 shows the thematic relationships in the INEX topics. Each topic title was stopped and stemmed; then if two topics shared a stem an edge was drawn between them (edgeless vertices were excluded). Vertex size is a function of topic length (in terms). The top left black topic, 585 (`international brigades spanish civil war`) shares terms with topics: 562 (`algerian war`); 618 (`lebanon militias war`); 553 (`spanish classical guitar players`); 615 (`spanish transition`); and 645 (`cellular phone international roaming`). A *war* theme is shared with two other topics and a *Spain* theme with two more. There are 71 edges in total.

If the number of edges (topics sharing a theme) is larger than that expected by chance then the topics are not independent. That is, a random selection of terms drawn from the Wikipedia vocabulary (of 2,012,641 terms) should result in a graph similar to Figure 4. To test this using the Bootstrap [5], 135 random queries matching the lengths shown in Figure 3 were generated and the number of links counted. Repeated a million times, it suggests that the probability of seeing 71 links is less than 0.01%. Thus thematic relationships are particular to the INEX topics; it is reasonable to conclude that FEATURE 5 does not hold.

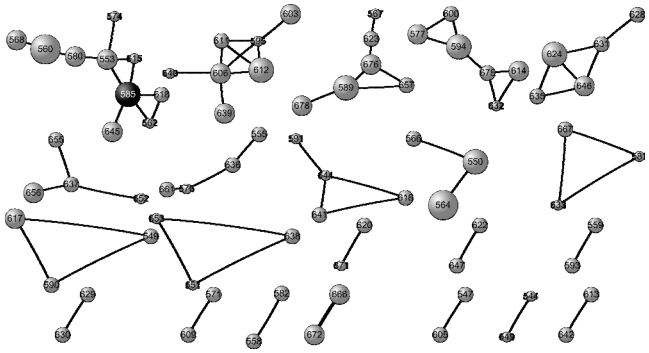

Figure 4: INEX 2008 topics sharing stemmed search terms

6. FULL RANGE OF NEEDS

The INEX topics should represent the full range of information needs expected of the Wikipedia. As an encyclopedia it is a collection of separate documents each on a single topic, a range of information needs might be selected by randomly choosing document titles from the collection, and a graph similar to Figure 4 would be expected.

Repeating the Bootstrap, but using document titles (not random queries) gives a probability of 0.6% of finding 71 edges. The expected number of edges for 135 titles is 28. The INEX topics are more tightly clustered than the collection. It is reasonable to conclude that FEATURE 4 is not satisfied.

7. DISCUSSION

Five features for robust evaluation are given and the INEX 2008 topic set evaluated against these. It is shown that the number of topics is sufficient, but that the full range of search engine features is not tested, they are not representative of user needs, do not represent the full range or needs, and are not independent.

If the INEX topic set is problematic then it is reasonable to expect an evaluative comparison of the two topic sets to show different results. Kamps *et al*. [3] compute Kendall's τ correlation of the relative rank order (using MAP) of all 163 runs submitted to the INEX 2008 ad hoc track against the 70 assessed INEX topics and 138 Proxy topics; indeed they weakly correlate (0.36). Our analysis questions the validity of the INEX topics and thus the results of INEX 2008.

REFERENCES

[1] Buckley, C. and E.M. Voorhees. *Evaluating evaluation measure stability*. SIGIR 2000, pp33-40.

[2] Hawking, D., et al., *Measuring Search Engine Quality.* Inf. Retr., 2001. **4**(1): 33-59.

[3] Kamps, J., M. Koolen, A. Trotman, *Comparative Analysis of Clicks and Judgments for IR Evaluation*, *WSCD 2009.*

[4] Lehtonen, M. *Designing User Studies for XML Retrieval*. SIGIR 2006 Workshop on XML Element Retrieval Methodology, pp28-34.

[5] Sakai, T., *Evaluating evaluation metrics based on the bootstrap*, SIGIR 2006, pp525-534.

[6] Trotman, A. and M. Lalmas. *Why Structural Hints in Queries do not Help XML Retrieval*. SIGIR 2006, pp711-712.