Let's Phrase It: INEX Topics Need Keyphrases

Antoine Doucet
GREYC CNRS UMR 6072
University of Caen Lower Normandy
F-14032 Caen Cedex
France
Antoine.Doucet @info.unicaen.fr

Miro Lehtonen
Department of Computer Science
P.O. Box 68
FI-00014 University of Helsinki
Finland
Miro.Lehtonen@cs.Helsinki.Fi

ABSTRACT

In this paper, we study and discuss the usage of phrases in the INEX evaluation of XML retrieval as well as in related research. We find that the INEX framework could easily become a unique testbed for researchers interested in the exploitation of complex terms in IR, while triggering interest from others. Unfortunately, our analysis of the use of keyphrases in INEX topics shows a downwards trend over the years that impacts on the attention of participants. While NEXI, the official query format of INEX, does indeed support keyphrases, its full potential does not materialize, as topic contents show a lack of consistency in their markup. In 2007, 87% of the INEX queries contained keyphrases, but only 11% of those were marked up. We present simple and low-cost solutions to let the INEX collections deliver their full potential in keyphrase retrieval.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Theory

Keywords: XML Information Retrieval, Phrase, Keyphrase

1. INTRODUCTION

Keyphrases are important for IR tasks. When used in queries they are matched with similar phrases in the documents. Advanced approaches are based on a keyphrase index which contains the most important if not all phrases of the document collection. INEX has provided excellent testbeds for keyphrase search. The document collections with marked up phrases and queries with explicitly marked keyphrases have inspired researchers to develop methods for parsing, indexing, and matching the phrases in order to improve from the basic keyword search. However, the annual evaluation is subject to criticism regarding keyphrase search. Even if a system with support for keyphrases outperforms the same system without keyphrases, we cannot conclude that the keyphrases actually improve the results as the number of queries involved is too low to make a statistically significant difference. Moreover, systems that can make the most out of keyphrases may not excel in the overall evaluation as they are not given any keyphrases for most of the queries.

Our analysis of the INEX data shows that keyphrases have been seriously neglected in the topic development of the recent INEX evaluations. Only 8% of the INEX 2007 topics define keyphrases, whereas the corresponding numbers were around 70% for the INEX 2003–2004 topics. In order to prevent keyphrases from going into complete oblivion, we suggest that the future topic authors of INEX be encouraged to mark keyphrases explicitly in the topic statements.

This paper is organized as follows. Essential previous research and statistics of phrase searching are presented in Section 2. Keyphrases in the INEX topics over the years are analysed in Section 3. Other connections between INEX and keyphrases are summarized in Section 4 including a brief overview on the document collections and methodology. In Section 5, we discuss the simple ways in which INEX could offer a unique testbed for the use of keyphrases in IR. We conclude the paper in Section 6.

2. KEYPHRASES AND IR TASKS

Numerous information retrieval systems rely on a "bag of words" representation of documents, and thus ignore the relative position of words. It is intuitively clear that taking phrases and collocations into account improves text understanding (for computer-based systems as well as for human readers). Zhai et al. [25] mention many subsequent problems. The biggest one is that of complex lexical units: The meaning of a word association is different from that of the "sum" of the meanings of the individual words they compose of. For instance, the expression "hot dog" is seldom used to refer to a dog. Another example is "to kick the bucket", an expression that means "to die", while "to kick" means "to strike out with the foot", and a "bucket" is a "cylindrical vessel used for holding or carrying liquids or solids". In such cases, it is clear that it is crucial to grasp the meaning of the expressions rather than solely that of the word components. Naturally, there have been numerous attempts to exploit lexical cohesion in IR.

2.1 Definition of a keyphrase

In this paper, we define a keyphrase as a set of adjacent terms, that are intended as a single lexical unit by the user. A keyphrase may be *explicit* (that is, clearly delimited with quotation marks or commas, for instance), or *implicit*, with no way to know, a priori, which sets of words the end user meant as phrases. Sample implicit and explicit keyphrases are shown in Table 1.

2.2 Previous research

Work on the use of phrases in IR has been undergone for over 30 years with mitigated success. Early results were very

SIGIR 2008 Workshop on Focused Retrieval
July 24, 2008, Singapore.
ACM .

promising. Unexpectedly, however, the constant growth of test collections caused a drastic fall in the quality of the results. In 1975, Salton et al. [16] showed an improvement in average precision over 10 recall points between 17% and 39% over a keyword-only baseline. In 1989, Fagan [3] reiterated the exact same experiments with a 10 Mb collection and obtained improvements from 11% to 20%. This negative impact of the collection size was lately confirmed by Mitra et al. [13] over a 655 Mb collection, improving the average precison by only one percent ! Turpin and Moffat [19] revisited and extended this work to obtain improvements between 4% and 6%.

A conclusion from this historical work is that keyphrases improve results at low levels of recall, but are globally inefficient for the *n* first ranked documents. According to Mitra et al. [13], such modest benefit from phrases to the best answers is explained by the fact that phrases promote documents that deal with only one aspect of possibly multi-faceted queries. For example, one of the TREC-4 topics is about “problems associated with pension plans, such as fraud, skimming, tapping or raiding”. Several top-ranked documents discuss pension plans, but not any related problem. Mitra et. al call this problem *inadequate query coverage*.

More recently, Vechtomova [20] started to investigate user-selected keyphrases, which puts aside the problem of phrase recognition so that we can focus on the usage of keyphrases in search tasks. The results show a consistent performance improvement in terms of average precision, and confirm the observation that the improvement is mainly built at low recall levels, while the impact is negative at higher levels.

2.3 Keyphrases in web search

An analysis of the query logs of the Excite search engine by Williams et al. [22] indicated that 5–10% of the web queries were phrase queries and that 41% of the rest also matched a phrase. In our terms, 5–10% of the queries are explicit keyphrases, and 41% of the rest may be implicit keyphrases. Unfortunately, commercial search engines are seldom using keyphrases per se, rather, they rely on proximity search or n-grams.

The generally recognized reason is that taking phrases into account is useful *in some cases*, while in others it only worsens retrieval performance. The key problem is that no way has yet been found to distinguish, *a priori*, the cases where keyphrases are useful from those where they are not. Therefore, it has been considered safer to rely on keywords and proximity search. From the current state of the art, it is straightforward to draw the conclusion that there is still plenty of room for future research on keyphrases in both web search and other search tasks.

3. KEYPHRASES IN INEX TOPICS

In this section, we study the usage of keyphrases in INEX topics over the 6 editions of the evaluation initiative (from 2002 to 2007). Two sample topics, respectively from INEX 2002 and INEX 2007 are presented in Figure 1 and Figure 2.

We have analysed all the accepted topics since 2002, and counted how many of them contain at least one implicit keyphrase and how many of them contain at least one explicit keyphrase. For this, we relied solely on the content of that XML element of the topic that was the most similar to a short, web-like, query. For the 2002–2004 campaigns, we used the <keywords> element, while for 2005 to 2007, we looked at the content of the <title> element.

Year	Topics	explicit KP	implicit KP	no KP
2002	60	20 (33%)	30 (75% *)	10 (17%)
2003	66	47 (71%)	13 (68% *)	6 (9%)
2004	74	51 (69%)	18 (78% *)	5 (7%)
2005	87	29 (33%)	52 (90% *)	6 (7%)
2006	125	43 (34%)	70 (85% *)	12 (10%)
2007	130	11 (8%)	102 (86% *)	17 (13%)
Total	542	201 (37%)	285 (84% *)	56 (10%)

Table 2: Number of accepted topics with explicit and implicit keyphrases. *) The percentage of implicit keyphrases is relative to the total number of topics without explicit keyphrases.

If the sequence of words was separated by some kind of delimiters (e.g., commas or quotations), and several words were found between those markers, we considered that the topic contained an explicit keyphrase. If the sequence of words contained no delimiters at all, and some adjacent words were clearly meant to be components of a complex lexical unit, the topic was considered to contain an implicit keyphrase. Typical examples are word pairs such as “information retrieval” or “firstname lastname”.

In Table 1, we describe how we judge the first 5 topics of INEX 2002. In topic 1, we can easily understand that “description logic” is meant to be a phrase, since the corresponding acronym, “DL” is also included. In topic 3, it is clear that “visualizing large information hierarchies” is an entity. We get a hint at this fact from the repetition of the word “information”, hence, “information spaces” is also intended as a phrase. In some cases, deciding whether or not a sequence of words was meant as an implicit keyphrase requires more consideration, but then, the topic description and narrative help us find the correct interpretation.

The per year statistics on the number of INEX topics containing explicit and implicit keyphrases are presented in Table 2.

We immediately notice that the proportion of topics containing keyphrases is much higher than that reported by Williams et al. [22] for web queries (37% of explicit keyphrases versus 5 to 10%, and 84% of implicit keyphrases amongst the rest instead of 41%). This is natural, as the INEX topics are much longer than web queries, and, unlike them, they were carefully thought up, reviewed, and selected by the organizers of the forum. This is actually one reason why it would take little additional effort to formalize keyphrase markup.

In 2003, the use of comma to separate entities in the <keywords> element was systematic. This certainly explains the surge in the ratio of explicit keyphrases. But the element name “keywords” was perhaps sometimes taken too literally, as in “keyword *versus* keyphrase”. Much to our surprise, several topics are using commas to separate words that are clear phrases. One of many examples is topic 101 shown in Figure 1, where it is very clear that “information retrieval” is a central concept, but the words “information” and “retrieval” are comma-separated in the <keywords> element!

The same phenomenon occurred in 2004, with numerous explicit keyphrases on one hand, and comma-separated keyphrase components on the other. The consistency of the comma-separation markup is additionally fading, as several

Topic	Word Sequence	explicit KP	implicit KP
1	description logic DL ABox TBox reasoning	no	yes
2	funding america DARPA US	no	no
3	visualizing large information hierarchies information spaces text multidimensional data datamining databases	no	yes
4	’extreme programming’ experiences results	yes	no
5	QBIC, IBM, image, video, content query, retrieval system	yes	no

Table 1: The “web-like queries” of the first 5 topics of INEX and our interpretation of whether a keyphrase is present or not.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_topic topic_id="101" query_type="CO" ct_no="37">
<title>+"t test" +information </title>
<description>use of the t-test in information retrieval </description>
<narrative>Information retrieval experimenters are advised to compare their new mean-average-precision
results with the baseline using a t test. We have reason to believe that this may be bad, even very bad,
advice, and are interested in papers that apply the t-tests to information retrieval (and possibly other
software engineering tasks). We are also interested in papers that mention alternative statistical
techniques to determine significance. </narrative>
<keywords>t-test, information, retrieval </keywords>
</inex_topic>
```

Figure 1: Topic 101, from INEX 2003.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_topic topic_id="522"  ct_no="190">
<title>April 19th revolution peaceful revolution velvet revolution quiet revolution</title>
<castitle>//*[about(.,"April 19th revolution" "peaceful revolution" "velvet revolution" "quiet revolution")]
</castitle>
<description>Find information about how the April 19th revolution differs from the peaceful,
velvet and quiet revolutions.</description>
<narrative>As a history buff, you have heard of the quiet revolution, the peaceful revolution and the
velvet revolution. For a skill-testing question to win an iPod you have been asked how they differ from
the April 19th revolution.</narrative>
</inex_topic>
```

Figure 2: Topic 522, from INEX 2007.

explicit phrases are double-marked with both commas and quotations marks, as in topic 158, where the `<keywords>` element contains:

`turing, test, consciousness, intelligence, "imitation game"`.

In 2005, there was no more element dedicated to keywords, which caused the beginning of a severe downwards trend on the number of explicit keyphrases. In 2006, an `<ontopic_keywords>` element was inluded in the topic format. Its influence is visible as the downwards trend was temporarily interrupted. Several participants have indeed included keyphrase delimiters in the `<ontopic_keywords>` element, that they then carried over to the `<title>` element. In 2007, unfortunately, the `<ontopic_keywords>` element disappeared, leading to a sharp fallout in terms of explicit keyphrases.

In conclusion, we make two major observations: 1) year after year, the number of explicit keyphrases has been decreasing, 2) the number of implicit keyphrases has been steady during the 6 INEX campaigns, notably, regardless of the topic format and guidelines.

Although it is harder to quantify, a third point should be made about the growing inconsistency of the phrase markup, whenever it was present. Some explicit keyphrases are marked with semi-columns, others with commas, and the rest with single or double quotation marks. Several keyphrases are double-marked (commas plus quotes).

It even seems that participants are confused about whether they are actually **allowed** to mark keyphrases. This impression is supported by examples such as topic 101, which we already mentioned (Figure 1), where two clear keyphrase components are comma-separated. Topic 522 is even more confusing (Figure 2), because the same keyword sequence is used in the `<castitle>` and in the `<title>` elements, with the exception that keyphrase delimitors are used in the `<castitle>` and removed in the `<title>` element. This is very hard to explain. Did the topic author have any reason to doubt that quotation marks were allowed inside `<title>` elements?

4. KEYPHRASES IN INEX RESEARCH

Keyphrases have been taken into account in various ways in the systems developed by INEX participants in the past years of INEX, e.g. when indexing and ranking the documents, or parsing the queries, or even generating formal queries from queries in a natural language. Examples of the different roles of keyphrases in INEX-related research are presented in this section.

4.1 XML documents

Any document with text in a natural language contains enough phrases to make it suitable for research on keyphrase search. From the keyphrase perspective, hypertext documents add an interesting feature to plain text documents: many phrases — the anchor texts — are now marked up with designated markers. HTML documents go even further down the road as they allow titles, emphasized content, list items etc. to be marked up with designated tags in addition to the anchor texts. However, HTML documents are computationally challenging to process for someone searching for phrases as the quality of the HTML code varies. In this sense, XML is the perfect document format for keyphrase search as the markup is highly regular and always well-formed[1].

The collections of XML documents provided by INEX 2002–2007 contain plenty of marked up phrases. In the collection of articles from IEEE journals, phrases are typically marked up because of an intended emphasis on the phrase, whereas the most common marked up phrases in the Wikipedia XML articles are anchor texts of hyperlinks. Both collections provide an interesting playground for methods on indexing and searching keyphrases. The methods that did not convince with plain text documents might have lots of unmaterialized potential with XML documents — if given a second chance.

4.2 Adhoc retrieval

Although some explicitly split the keyphrases of INEX queries into unordered sets of individual keywords [1], many others parse them and match them with similar phrases in the documents. Both Raja et al. and Lehtonen and Doucet compute a separate score for keyphrase similarity in the vector space model which is combined with a keyword similarity score into a single Retrieval Status Value [15, 12]. A less strict interpretation of the concept of a keyphrase is defined in the TRIX system where the words of the keyphrase are only required to appear in a mutual XML element [8].

Approaches where the documents are stored in an XML database [14, 21] inherently support constraints on the order and distance of the keywords. Other database systems supporting keyphrase queries include Cheshire II with proximity indexes [11] as well as TopX with term offsets stored in an auxiliary database table [17].

Phrases are also part of the document representation in several systems for XML IR, e.g. Maximal Frequent Sequences of EXTIRP [2], Rich Document Representation of the extended PLIR [10], and bigram language models of TIJAH [9]. Despite the quite widespread interest in parsing and matching keyphrases, most of the efforts originate in the early years of INEX. The obvious explanation for the recent lack of interest lies in the diminishing proportion of explicitly marked keyphrases in the INEX topics.

4.3 The NLP Track prospect

Besides the adhoc track queries, keyphrases have had an important role in the experiments of the NLP track of INEX. The main challenge of the track has been to convert a natural language query into a formal NEXI query which is the official query format of INEX [18]. One of the earliest systems participating in the track was NLPX which segments sentences into disjoint chunks that eventually converge into NEXI keyphrases [23, 24]. Zargayouna et al. go along the same lines as they *"prefer complex terms to simple ones"* as they generate NEXI queries [7]. However unfortunate it may seem that the official NEXI queries created by INEX participants come with so few explicitly marked keyphrases, the NLP track may regard this as an opportunity to assist by applying various chunking methods to the topic titles. The other option is to have the keyphrases marked by topic developers, but few participants are currently doing it.

5. DISCUSSION

In information retrieval, the use of phrases includes two problems: the first one is that of the detection of phrases

[1] Other than well-formed XML is not defined.

in the document collection, while the second one is to find ways to improve retrieval effectiveness, given good phrases.

As we have seen in Section 4.1, XML mark-up has potential to ease the extraction of phrases, which is a way to isolate the problem of phrase exploitation, that no other evaluation framework provides, even when they show specific consideration for phrase usage. For instance, the topics of NTCIR collections[2], have systematically included a `<CONC>` element, with a list of keywords in which keyphrases are clearly and consistently delimited, but the documents themselves have no comparable phrase-related markup.

As opposed to most IR evaluation forums, we must point out a specificity of INEX, which is that topics are collaboratively contributed by participants. The plentiful of authors makes it naturally harder to keep consistent notation across the topic set, especially when keyphrase markup has barely ever been mentioned in the topic creation guidelines.

Unfortunately, while the XML mark-up of the INEX collections permits to ease phrase extraction from documents, keyphrases have seemingly been abandoned in the topics. This has naturally led to a drop in their usage by INEX participants. As detailed in Section 4, while several made use of complex terms in the first years, when more explicit phrases were available, the interest in keyphrases has lately diminished drastically: Most new participants are not taking phrases into account.

An easy way to solve the issue mentioned here is to define a strict way to mark keyphrases, and request topics to conform to it. The goal would be to replace all the implicit keyphrases with explicit ones, which could even be corrected by the organizers, although leaving it to the participants is certainly preferable. Clearly, the additional amount of work is small, nearly negligible.

Having a large set of clearly marked up phrases, together with the implicit phrases contained in the document collection should make INEX a unique testbed for researchers interested in the use of complex terms in IR. At the same time, it makes no harm to others, who can very easily ignore the keyphrase delimitors.

However, we must point out that our goal is not only to please participants interested in the use of keyphrases, but to avoid causing any inconvenience to others. We hope that the generalization of explicit keyphrases will encourage all participants to get involved, which should come naturally when statistically siginficant performance improvements are achieved (retrieval performance has reportedly been improved, but the small number of keyphrases does not leave a chance to statistically significant results without a massive performance surge).

6. CONCLUSION

The topic definitions of the past INEX evaluations have had several shortcomings regarding the formulation of keyphrases. An INEX evaluation of keyphrase queries is currently a lost opportunity despite its potential. Nonetheless, we are not far from solving the major problems and seeing a brighter future for the research on searching XML documents for keyphrases. The key facts that we have shown in this paper are the following:

- Keyphrases are common in real world queries, but the methodology for processing them is not yet mature.
- INEX document collections are rich in phrases, but regarding the INEX topics, keyphrases are about to become extinct.
- Several systems with result submissions to the past INEX evaluations support keyphrase queries.

The only thing needed for a decent INEX evaluation of keyphrase search are more queries — INEX topics — with explicitly marked keyphrases. The additional effort required is minimal but invaluable as it has the potential to revive the research activity in the area of keyphrase search. The first step entails writing more detailed guidelines for topic development, and the second step, collective topic authoring by the INEX participants.

7. REFERENCES

[1] C. L. Clarke and P. L. Tilker. Multitext experiments for inex 2004. In Fuhr et al. [5], pages 85–87.

[2] A. Doucet, L. Aunimo, M. Lehtonen, and R. Petit. Accurate Retrieval of XML Document Fragments using EXTIRP. In *INEX 2003 Workshop Proceedings*, pages 73–80, Schloss Dagstuhl, Germany, 2003.

[3] J. L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40:115–132, 1989.

[4] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors. *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 2005, Revised Selected Papers*, volume 3977 of *Lecture Notes in Computer Science*. Springer, 2006.

[5] N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors. *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 2004*, volume 3493 of *Lecture Notes in Computer Science*. Springer, 2005.

[6] N. Fuhr, M. Lalmas, and A. Trotman, editors. *Comparative Evaluation of XML Information Retrieval Systems, 5th International workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Dagstuhl Castle, Germany, December 2006*, volume 4518 of *Lecture Notes in Computer Science*. Springer, 2007.

[7] Haïfa Zargayouna and Victor Rosas and Sylvie Salotti. Shallow Parsing of INEX Queries. In Fuhr et al. [6], pages 284–293.

[8] Jaana Kekäläinen and Marko Junkkari and Paavo Arvola and Timo Aalto. TRIX 2004 — struggling with the overlap. In Fuhr et al. [5], pages 127–137.

[9] Johan List and Vojkan Mihajlović and Georgina Ramírez and Arjen de Vries and Djoerd Hiemstra and Henk Ernst Blok. TIJAH: Embracing IR Methods in XML Databases. *Information Retrieval*, 8(4):547–570, 2005.

[10] M. Karimzadegan, J. Habibi, and F. Oroumchian. Logic-based XML information retrieval for determining the best element to retrieve. In Fuhr et al. [5], pages 88–99.

[2] "NII Test Collection for IR systems", http://research.nii.ac.jp/~ntcadm/index-en.html

[11] R. R. Larson. Cheshire II at INEX '04: Fusion and Feedback for the Adhoc and Heterogeneous Tracks. In Fuhr et al. [5], pages 322–336.

[12] M. Lehtonen and A. Doucet. Extirp: Baseline retrieval from wikipedia. In Fuhr et al. [6], pages 119–124.

[13] M. Mitra, C. Buckley, S. A., and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO97, Computer-Assisted Information Searching on the Internet*, pages 200–214, 1997.

[14] J. Pehcevski, J. Thom, and A.-M. Vercoustre. Combining information retrieval and a native XML database. *Information Retrieval*, 8(4):571–600, 2005.

[15] F. Raja, M. Keikha, M. Rahgozar, and F. Oroumchian. Using rich document representation in XML information retrieval. In Fuhr et al. [6], pages 294–301.

[16] G. Salton, C. Yang, and C. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26:33–44, 1975.

[17] M. Theobald, R. Schenkel, and G. Weikum. TopX and XXL at INEX 2005. In Fuhr et al. [4], pages 282–295.

[18] Trotman,Andrew and Sigurbjörnsson,Börkur. Narrowed extended xpath i (nexi). In Fuhr et al. [5], pages 16–40.

[19] A. Turpin and A. Moffat. Statistical phrases for vector-space information retrieval. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 309–310, 1999.

[20] O. Vechtomova. The role of multi-word units in interactive information retrieval. In *Proceedings of the 27th European Conference on Information Retrieval, Santiago de Compostela, Spain*, pages 403–420, 2005.

[21] F. Weigel, K. U. Schulaz, and H. Meuss. Ranked retrieval of structured documents with the S-Term vector space model. In Fuhr et al. [5], pages 238–252.

[22] H. E. Williams, J. Zobel, and D. Bahle. Fast phrase querying with combined indexes. *ACM Transactions on Information Systems (TOIS)*, 22(4):573–594, 2004.

[23] A. Woodley and S. Geva. NLPX at INEX 2005. In Fuhr et al. [4], pages 358–372.

[24] A. Woodley and S. Geva. NLPX at INEX 2006. In Fuhr et al. [6], pages 302–311.

[25] Zhai, Chengxiang, X. Tong, N. Milic Frayling, and E. D.A. Evaluation of syntactic phrase indexing. In *Proceedings of the 5th Text Retrieval Conference, TREC-5*, pages 347–358, 1997.