Presenting Structured Text Retrieval Results

Jaap Kamps
University of Amsterdam
http://staff.science.uva.nl/~kamps/

SYNONYMS

None

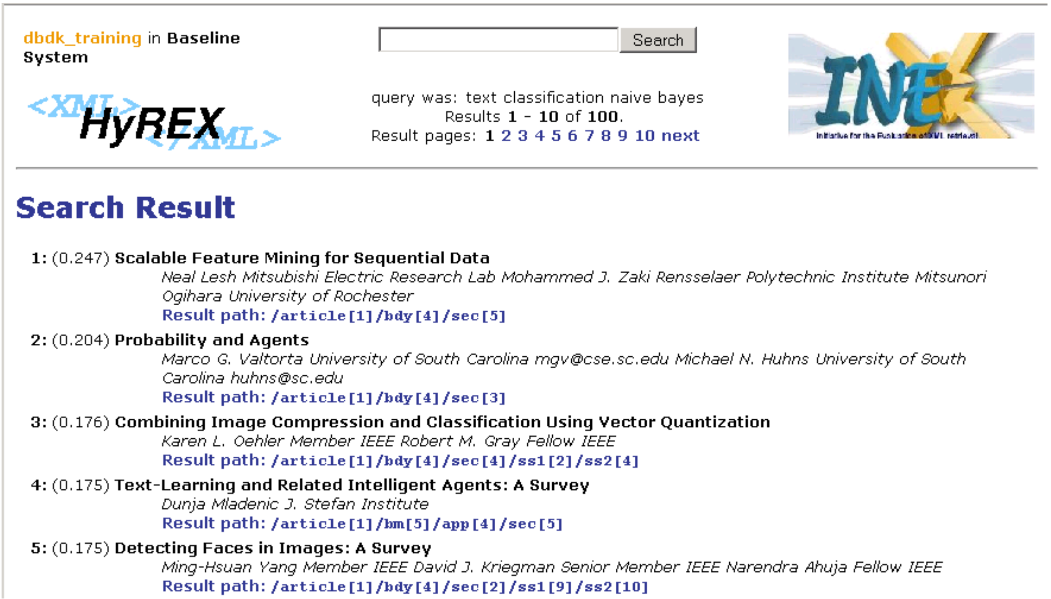
DEFINITION

Presenting structured text retrieval results refers to the fact that, in structured text retrieval, results are not independent and a judgment on their relevance needs to take their presentation into account. For example, HTML/XML/SGML documents contain a range of nested sub-trees that are fully contained in their ancestor elements. As a result, structured text retrieval should make explicit the assumptions on how the retrieval results are to be presented. Four of the main assumptions to be addressed are the following. First, the *unit of retrieval* assumption: is there a designated retrieval unit (such as the document or root node of the structured document) or can every sub-tree be retrieved in principle? Second, the *overlap* assumption: may retrieval results contain text or content already part of other retrieval results (such as a full article and one of its individual paragraphs)? Third, the *context* assumption: can results from the same structured document be interleaved with results from other structured documents? Fourth, the *display* assumption: is a retrieval result (say a document sub-tree corresponding to a paragraph) presented as an autonomous unit of text, or as an entry-point within a structured document?

HISTORICAL BACKGROUND

Although similar considerations play an important role in the design of user interfaces (see, for example, Hearst [6]), this entry will focus on the underlying principles of the different structured text retrieval tasks. Structured text retrieval dates back, at least, to the early days of passage retrieval [13]. Early passage retrieval approaches have been using either the document structure (sentences, paragraphs, sections, etc.), or arbitrary text windows of fixed length. The early experimental results primarily confirmed the effectiveness of passage-level evidence for boosting document retrieval. The use of document structure derived from SGML mark-up was pioneered by Wilkinson [20], studying adhoc SGML element retrieval. Probabilistic indexing approaches for databases have been studied even earlier [4], allowing to rank results based on vague queries. Adhoc XML element retrieval and best entry point retrieval was studied in the Focus project [3, 8].

The main thrust in recent years is the initiative for the evaluation of XML retrieval INEX [7]. The retrieval task descriptions heavily evolved during the different years. Initially, in 2002, INEX studied adhoc XML element retrieval for keyword (Content-Only) and structured (Content-And-Structure) queries with the goal to "retrieve the most specific relevant document components" [5, p.2]. This generic adhoc XML element retrieval task was continued at INEX 2003 [9, p.200] and at INEX 2004 [11, p.237], asking for "components that are most specific and most exhaustive with respect to the topic of request." Ongoing discussion, and vivid disagreement, on the interpretation of generic adhoc XML element retrieval task prompted the introduction of three different retrieval strategies at INEX 2005 [10, p.385-386]: *Thorough* aims to find all highly exhaustive and specific elements (roughly corresponding to the earlier INEX task); *Focussed* aims to find the most specific and exhaustive element in path (no overlapping results); and *Fetch and browse* aims to first identify relevant articles, and then to find the most specific and exhaustive elements within the fetched articles (results grouped by article). These different adhoc XML element retrieval tasks have been continued and further explicated at INEX 2006 [1], with the Fetch and



dbdk_training in **Baseline System**

query was: text classification naive bayes
Results **1** - **10** of **100**.
Result pages: **1** 2 3 4 5 6 7 8 9 10 next

Search Result

1: (0.247) **Scalable Feature Mining for Sequential Data**
Neal Lesh Mitsubishi Electric Research Lab Mohammed J. Zaki Rensselaer Polytechnic Institute Mitsunori Ogihara University of Rochester
Result path: /article[1]/bdy[4]/sec[5]

2: (0.204) **Probability and Agents**
Marco G. Valtorta University of South Carolina mgv@cse.sc.edu Michael N. Huhns University of South Carolina huhns@sc.edu
Result path: /article[1]/bdy[4]/sec[3]

3: (0.176) **Combining Image Compression and Classification Using Vector Quantization**
Karen L. Oehler Member IEEE Robert M. Gray Fellow IEEE
Result path: /article[1]/bdy[4]/sec[4]/ss1[2]/ss2[4]

4: (0.175) **Text-Learning and Related Intelligent Agents: A Survey**
Dunja Mladenic J. Stefan Institute
Result path: /article[1]/bm[5]/app[4]/sec[5]

5: (0.175) **Detecting Faces in Images: A Survey**
Ming-Hsuan Yang Member IEEE David J. Kriegman Senior Member IEEE Narendra Ahuja Fellow IEEE
Result path: /article[1]/bdy[4]/sec[2]/ss1[9]/ss2[10]

Figure 1: Displaying structured text retrieval results as a ranked list of elements (reproduced from [12]).

browse task refined to: *Relevant in Context* aims to retrieve a set of non-overlapping relevant elements per article; and *Best in Context* aims to retrieve, per article, a single best entry point to read its relevant content. At INEX 2007 three tasks are continued: Focused, Relevant in Context, and Best in Context, but liberalized to arbitrary passages [2].

SCIENTIFIC FUNDAMENTALS

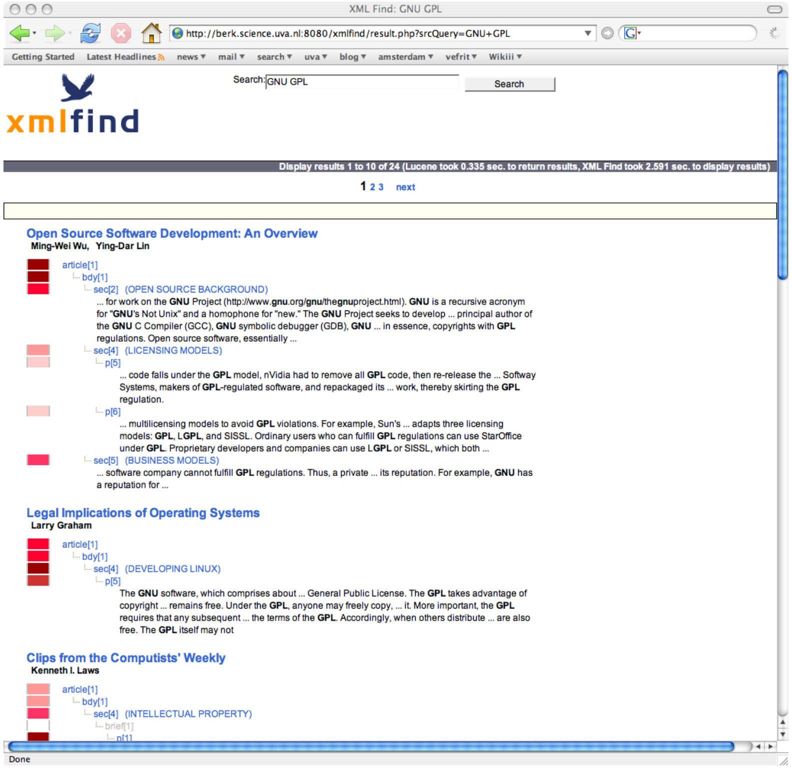
The way in which retrieval results are presented to users, is always a crucial factor determining the success or failure of an operational retrieval system. However, within the Cranfield/TREC tradition of evaluating document retrieval systems it is unproblematic to abstract away from presentation issues and analyze retrieval effectiveness by regarding retrieved documents as atomic and independent results. In structured text retrieval, the situation is different, and there is a need to make explicit some of the assumptions underlying the retrieval task since these have an impact on what is regarded as a "relevant" retrieval result.

First, the *unit of retrieval* assumption: is there a designated retrieval unit (such as the document or root node of the structured document) or can every sub-tree be retrieved in principle? Rather than treating documents as atomic, structured documents have internal document structure that allows any logical unit of them to be retrieved. For example, in case of a textual document where the layout structure is marked up, it is possible to retrieve sections, paragraphs, or still the whole article if its completely devoted to the topic of request. Figure 1 contains a screen-shot of a XML element retrieval system that retrieves a ranked list of XML elements.

Second, the *overlap* assumption: may retrieval results contain text or content already part of other retrieval results (such as a full article and one of its individual paragraphs)? Interactive experiments at INEX 2004 [17] clearly revealed that test persons disliked a ranked list of element results that overlap in whole or part in their content. Hence, if the retrieval tasks should reflect a scenario in which the ranked elements are directly displayed to an end-user, retrieval results should be disjoint.

Third, the *context* assumption: can results from the same structured document be interleaved with results from other structured documents? A further finding of the interactive experiments at INEX 2004 [17] is that test persons prefer results from the same document be grouped together. Figure 2 contains a screen-shot of a XML element retrieval system that retrieves XML elements displayed in document order in their article's context.

Fourth, the *display* assumption: is a retrieval result (say a document sub-tree corresponding to a paragraph)



Display results 1 to 10 of 24 (Lucene took 0.335 sec. to return results, XML Find took 2.591 sec. to display results)

Open Source Software Development: An Overview
Ming-Wei Wu, Ying-Dar Lin

🟥 article[1]
🟥 bdy[1]
🟥 sec[2] (OPEN SOURCE BACKGROUND)
... for work on the **GNU** Project (http://www.**gnu**.org/**gnu**/the**gnu**project.html). **GNU** is a recursive acronym for "**GNU**'s Not Unix" and a homophone for "new." The **GNU** Project seeks to develop ... principal author of the **GNU** C Compiler (GCC), **GNU** symbolic debugger (GDB), **GNU** ... in essence, copyrights with **GPL** regulations. Open source software, essentially ...
🟥 sec[4] (LICENSING MODELS)
🟥 p[5]
... code falls under the **GPL** model, nVidia had to remove all **GPL** code, then re-release the ... Softway Systems, makers of **GPL**-regulated software, and repackaged its ... work, thereby skirting the **GPL** regulation.
🟥 p[6]
... multilicensing models to avoid **GPL** violations. For example, Sun's ... adapts three licensing models: **GPL**, L**GPL**, and SISSL. Ordinary users who can fulfill **GPL** regulations can use StarOffice under **GPL**. Proprietary developers and companies can use L**GPL** or SISSL, which both ...
🟥 sec[5] (BUSINESS MODELS)
... software company cannot fulfill **GPL** regulations. Thus, a private ... its reputation. For example, **GNU** has a reputation for ...

Legal Implications of Operating Systems
Larry Graham

🟥 article[1]
🟥 bdy[1]
🟥 sec[4] (DEVELOPING LINUX)
🟥 p[5]
The **GNU** software, which comprises about ... General Public License. The **GPL** takes advantage of copyright ... remains free. Under the **GPL**, anyone may freely copy, ... it. More important, the **GPL** requires that any subsequent ... the terms of the **GPL**. Accordingly, when others distribute ... are also free. The **GPL** itself may not

Clips from the Computists' Weekly
Kenneth I. Laws

🟥 article[1]
🟥 bdy[1]
🟥 sec[4] (INTELLECTUAL PROPERTY)
□ brief[1]
🟥 p[1]

Figure 2: Displaying structured text retrieval results within article context (reproduced from [14]).

Table 1: structured text retrieval tasks.

	Unit of Retrieval	Overlap	Context	Display
Article retrieval	Whole article	–	–	–
Thorough	Arbitrary element	Allowed	Scattered	Elements
Focussed/Focused	Arbitrary element	Non-overlapping	Scattered	Elements/Passages
Fetch and browse	Arbitrary element	Allowed	List per article	Elements
Relevant in Context	Arbitrary element	Non-overlapping	Set per article	Elements/Passages
Best in Context	Arbitrary element	Non-overlapping	One result per article	Starting point

presented as an autonomous unit of text, or as an entry-point within a structured document? A decision on the relevance of a particular document component crucially depends on whether it will be presented as an isolated excerpt, or within its original context. In the first case, the component should be fully self-contained: it should not only contain the relevant information (say, for example, a description of an algorithm) but also establish that this information is, indeed, satisfying the topic of request (for example, that the algorithm is the fastest way to lexicographically sort a list, if that were the topic of request). This is related to linguistics, where there is a common distinction between the context (or topic/theme: that what is being talked about), and the information (or comment/rheme/focus: that what is being said). If results are to be presented in their document context, the link to the topic of request can be taken for granted and only the sought information can be regarded as relevant. If results are to be presented out of context, both the information and its relation to the topic of request are needed to establish the relevance of a document component.

Table 1 shows how the different structured text retrieval tasks are based on different underlying assumptions. For traditional document or article retrieval, there is a fixed unit of retrieval and assumptions on overlap, context, or display do not apply. For the generic adhoc element retrieval task (INEX 2002-2004) or Thorough (INEX 2005-2006), any document component can be retrieved, and there are no restrictions on overlap, context, or display. Basically, the task is system-biased, reflecting the ability of the retrieval engine to estimate the relevance of individual document components, for example for further processing methods. For Focussed/Focused (INEX 2005-2007), a ranked list of non-overlapping document components is asked for, with no restrictions on context or display. This task reflects a scenario where a ranked-list of document components is directly presented to the searcher. For Fetch and browse (INEX 2005), retrieval results from the same structured document need to be returned consecutive, with no restriction on overlap or display. This results in a tasks resembling on the one hand traditional document retrieval, whilst on the other hand providing deep-linking to relevant document components. The same holds for Relevant in context (INEX 2006-2007), where there is an unranked set of now non-overlapping elements per article, reflecting results to be presented in document order. Finally, Best in context (INEX 2006-2007) explicitly asks for a single best entry point into the article (so non-overlapping and non-scattered articles by definition). This scenario captures a "relative" notion of relevance, where users desire access to the best information, rather than all relevant information.

These different retrieval tasks lead to different evaluations of what systems and techniques are effective for structured text retrieval. Although these tasks are not unrelated, for example, the generic Thorough task (capturing the ability of a system to estimate the relevance of an element) can be use as input for further processing for the other tasks, each of these different retrieval tasks is capturing a different aspect of structured text. The retrieval tasks bring in elements from the task context in which they are to be applied, either in a end-user setting or system setting. As a result, the richer descriptions of the task's context and underlying assumptions are resonating more closely with actual real-world applications [19]. Bringing task-specific elements into information retrieval benchmark testing has been identified as one of the main research directions for further enhancing information access in general [15].

KEY APPLICATIONS

Structured text retrieval has the potential to improve information access by giving more direct access to the relevant information inside documents. As Salton et al. [13, p.49] put it:

> Large collections of full-text documents are now commonly used in automated information retrieval. When the stored document texts are long, the retrieval of complete documents may not be in the users' best interest. In such circumstances, efficient and effective retrieval results may be obtained by using passage retrieval strategies designed to retrieve text excerpts of varying size in response to statements of user interest.

Structured document retrieval is becoming increasingly important in all areas of information retrieval, the application to full-text book searching is obvious and such commercial systems already exist [19].

FUTURE DIRECTIONS

Improving information access by formulating retrieval tasks that capture interesting aspects of real-world structured text searching is an ongoing open problem. There has been a series of workshops addressing open problems, including real-world applications, the unit of retrieval, tasks and measures, and the problem of overlap [18].

The traditional picture of IR takes as input a document collection and a query, and gives as output a ranked list of documents. In the retrieval task, there no distinction between the hit list (communicating the ranked list) and the actual result documents. Where structured document retrieval is going beyond a linear ranked list of results, at least conceptually, interesting new research questions present themselves. By presenting related results from the same article, like in Figure 2, the hit-list becomes a query-biased summary of the discourse structure of the retrieved article. Szlávik et al. [16] conduct experiments on the level of detail desired by searchers. Evaluation of such a system seems to require taking both retrieval effectiveness and document summarization aspects into account.

DATA SETS

Notable data-sets are:
The *Shakespeare test collection* used in the Focus project 2000–2001 [3].
The *IEEE Computer Society collection* used at INEX 2002–2004 [7].
The expanded *IEEE Computer Society collection* used at INEX 2005 [7].
The *Wikipedia XML Corpus* used at INEX 2006–2007 [7].

CROSS REFERENCE

Evaluation metrics for structured text retrieval
INitiative for the Evaluation of XML Retrieval (INEX)
XML Retrieval

RECOMMENDED READING

[1] C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *INEX 2006 Workshop Pre-Proceedings*, pages 381–388, 2006.

[2] C. L. A. Clarke, J. Kamps, and M. Lalmas. INEX 2007 retrieval task and result submission format. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Pre-Proceedings of INEX 2007*, pages 445–453, 2007.

[3] Focus. Focussed retrieval of structured documents – a large experimental study, 2001. `http://qmir.dcs.qmul.ac.uk/Focus/index.htm`.

[4] N. Fuhr. A probabilistic framework for vague queries and imprecise information in databases. In D. McLeod, R. Sacks-Davis, and H.-J. Schek, editors, *16th International Conference on Very Large Data Bases*, pages 696–707. Morgan Kaufmann, 1990.

[5] N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors, *Proceedings of the First Workshop of the INitiative for the Evaluation of XML retrieval (INEX)*, pages 1–17. ERCIM Publications, 2003.

[6] M. A. Hearst. User interfaces and visualization. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, chapter 10, pages 257–324. ACM Press, New York and Addison Wesley Longman, Harlow, 1999.

[7] INEX. INitiative for the Evaluation of XML Retrieval, 2007. `http://inex.is.informatik.uni-duisburg.de/`.

[8] G. Kazai, M. Lalmas, and J. Reid. Construction of a test collection for the focussed retrieval of structured documents. In F. Sebastiani, editor, *Advances in Information Retrieval, 25th European Conference on IR Research (ECIR 2003)*, volume 2633 of *Lecture Notes in Computer Science*, pages 88–103. Springer, 2003.

[9] G. Kazai, M. Lalmas, N. Gövert, and S. Malik. INEX'03 retrieval task and result submission specification. In *INEX 2003 Workshop Proceedings*, pages 200–203, 2004.

[10] M. Lalmas and G. Kazai. INEX 2005 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *INEX 2005 Workshop Pre-Proceedings*, pages 385–390, 2005.

[11] M. Lalmas and S. Malik. INEX 2004 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *INEX 2004 Workshop Pre-Proceedings*, pages 237–240, 2004.

[12] S. Malik, C.-P. Klas, N. Fuhr, B. Larsen, and A. Tombros. Designing a user interface for interactive retrieval of structured documents – lessons learned from the INEX interactive track. In *10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 291–302, 2006.

[13] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58. ACM Press, New York NY, 1993.

[14] B. Sigurbjörnsson. *Focused Information Access using XML Element Retrieval.* SIKS dissertation series 2006-28, University of Amsterdam, 2006.

[15] K. Sparck Jones. What's the value of TREC – is there a gap to jump or a chasm to bridge? *SIGIR Forum*, 40(1):10–20, 2006.

[16] Z. Szlávik, A. Tombros, and M. Lalmas. Feature- and query-based table of contents generation for XML documents. In *29th European Conference on Information Retrieval (ECIR)*, pages 456–467, 2007.

[17] A. Tombros, B. Larsen, and S. Malik. The interactive track at INEX 2004. In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *Advances in XML Information Retrieval. Third Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, pages 410–423. Springer Verlag, Heidelberg, 2005.

[18] A. Trotman, S. Geva, and J. Kamps, editors. *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 2007. University of Otago, Dunedin New Zealand.

[19] A. Trotman, N. Pharo, and M. Lehtonen. XML-IR users and use cases. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Comparative Evaluation of XML Information Retrieval Systems: Fifth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2006)*, volume 4518 of *Lecture Notes in Computer Science*, pages 400–412. Springer Verlag, Heidelberg, 2007.

[20] R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 311–317. Springer-Verlag, New York NY, 1994.