Term statistics for structured text retrieval

Mounia Lalmas, Department of Computer Science, Queen Mary, University of London

SYNONYM
> Within-element term frequency, Inverse element frequency

DEFINITION
Classical ranking algorithms in information retrieval make use of term statistics, the most common (and basic) ones being within-document term frequency, *tf*, and document frequency, *df*. *tf* is the number of occurrences of a term in a document and is used to reflect how well a term captures the topic of a document, whereas *df* is the number of documents in which a term appears and is used to reflect how well a term discriminates between relevant and non-relevant documents. *df* is also commonly referred to as inverse document frequency, *idf*, since it is inversely related to the importance of a term. Both *tf* and *idf* are obtained at indexing time. Ranking algorithms for structured text retrieval, and more precisely XML retrieval, require similar terms statistics, but with respect to elements.

MAIN TEXT
To calculate term statistics for elements, one could simply replace documents by elements and calculate so-called within-element term frequency, *etf*, and inverse element frequency, *ief*. This however raises an issue because of the nested nature of, in particular, XML documents. For instance, suppose that a section element is composed of two paragraph elements. The fact that a term appears in the paragraph necessitates that it also appears in the section. This overlap can be taken into account when calculating the *ief* value of a term.

In structured retrieval, in contrast to "flat" document retrieval, there are no a priori fixed retrieval units. The whole document, a part of it (e.g. one of its section), or a part of a part (e.g. a paragraph in the section), all constitute potential answers to queries. The simplest approach to allow the retrieval of elements at any level of granularity is to index all elements. Each element thus corresponds to a document, and *etf* and *ief* for each element are calculated based on the concatenation of the text of the element and that of its descendants (e.g. [4]).

With respect to the calculation of the inverse element frequency, *ief*, the above approach ignores the issue of nested elements. Indeed, the *ief* value of a term will consider both the element that contains that term and all elements that do so in virtue of being ancestors of that element. Alternatively, *ief* can be estimated across elements of the same type (e.g. [3]) or across documents (e.g. [1]). The former greatly reduces the impact of nested elements on the *ief* value of a term, but does not eliminate it as elements of the same type can be nested within each other. This approach can be extended to consider the actual path of an element, leading to so-called inverted path frequency. For example, in [2], this is defined as the combination of the *ief* values (as above calculated) with respect to each of the element types forming the path. The latter case, i.e. calculating *ief* across documents, is the same as using inverse document frequency, which completely eliminates the effect of nested elements.

CROSS REFERENCES
XML Retrieval
Indexing units
Structure weight
Relationships in structured text retrieval

RECOMMENDED READING

[1] Clarke, C. L. A. (2005). Controlling Overlap in Content-Oriented XML Retrieval. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil*, pp 441-448.

[2] Grabs, G., & Schek, H.-S. (2002). ETH Zürich at INEX: Flexible Information Retrieval from XML with PowerDB-XML. *Proceedings of the First INEX Workshop (INEX 2002), Dagstuhl, Germany, ERCIM*, pp 141-148.

[3] Mass, Y., & Mandelbrod, M. (2005). Component ranking and automatic query refinement for XML retrieval. *Advances in XML Information Retrieval and Evaluation*, *Proceedings of the Fourth INEX Workshop (INEX 2005), Dagstuhl, Germany, Lecture Notes in Computer Science 3493*, pp 73-84.

[4] Sigurbjörnsson, B., Kamps J., & de Rijke, M. (2003). An element-based approach to XML retrieval. *Proceedings of the Second INEX Workshop (NEX 2003), Dagstuhl, Germany*, pp, 19-26. Available at: http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf.