Structure weight

Mounia Lalmas, Department of Computer Science, Queen Mary, University of London, mounia@acm.org

SYNONYM

None

DEFINITION

In structured text retrieval, the structure of a text component may be used to estimate the relevance of that component. This is done by associating a weight to the structure reflecting its significance when estimating the relevance of the component for a given query.

MAIN TEXT

Associating weight to the structure of a component in itself is not new, and several investigations have been reported for whole document retrieval. This entry is concerned with structure weights in the context of structured text retrieval, where the aim is to exploit the document structure to return document components, instead of whole documents.

In structured text retrieval, not all document components will trigger the same user satisfaction when returned as answers to queries. In the context of structured documents mark-up in XML, some document components, i.e. XML elements, may not be appropriate to return because they are too small, of a tag type that does not contain informative content, nested too deep in the document logical structure, or for other reasons. When ranking XML elements, their structure (size, tag type, path, depth, etc.) may prove important. The importance of the element structure is captured through a weight, which can be binary.

Using binary weights means that an element is (value one) or is not (value zero) considered for indexing and retrieval. The decision can be made by looking at the DTD[1] of the collection, past relevance data, and/or the requirements of the application and user scenario. In the selective indexing strategy [3], only elements of types that were found to contain relevant content for previous query sets (relevance data) are considered. Any elements with a length size less than a given threshold can also be ignored.

Weights can be assigned to characteristics of elements, such as length, depth, location in the document logical structure, and so on. For instance, within the language modelling framework, length has been used as a normalization parameter (weight) incorporated through a prior probability in the ranking formula [2].

With statistical approaches, the weights are estimated based on training data, such as past relevance data. The weights can be determined using machine learning, and then used in the ranking function. They can also be directly calculated based on the

[1] Document Type Definition.

distribution of element characteristics. For example, in [1], the distribution of tag types is used in a way similar to the binary independence retrieval model (investigating the "presence" of tags in relevant and non-relevant elements) to estimate the element weights.

CROSS REFERENCES

Indexing units
Logical structure
Relationships in structured retrieval
XML Retrieval

RECOMMENDED READING

[1] M. Gery, C. Largeron and F. Thollard. Probabilistic document model integrating XML structure. INEX 2007 Pre-proceedings, pp 139–149, 2007. http://inex.is.informatik.uni-duisburg.de/2007/inex07/pdf/2007-preproceedings.pdf.

[2] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval, *ACM SIGIR Conference on Research and Development in Information Retrieval*, *Sheffield, UK*, pp 80– 87, 2004.

[3] Y. Mass and M. Mandelbrod. Component ranking and automatic query refinement for XML retrieval. *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*, pp 73-84, 2005.