Abstract

Query Expansion using Pseudo Relevance Feedback is a useful and a popular technique for reformulating the query. In our proposed query expansion method, we assume that relevant information can be found within a document near the central idea. The document is normally divided into sections, paragraphs and lines. The proposed method tries to extract keywords that are closer to the central theme of the document. The expansion terms are obtained by equi-frequency partition of the documents obtained from pseudo relevance feedback and by using tf-idf scores. The idf factor is calculated for number of partitions in documents. The group of words for query expansion is selected using the following approaches: the highest score, average score and a group of words that has maximum number of keywords. As each query behaved differently for different methods, the effect of these methods in selecting the words for query expansion is investigated. From this initial study, we extend the experiment to develop a rule-based statistical model that automatically selects the best group of words incorporating the tf-idf scoring and the 3 approaches explained here, in the future. The experiments were performed on FIRE 2011 Adhoc Hindi and English test collections on 50 queries each, using Terrier as retrieval engine.
Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval

References

- Jones R, Rey B, Madani O and Greiner W. Generating Query Substitutions. In Proceedings of the 15th International Conference on World Wide Web. ACM 2006, pp. 387-396.
- Croft WB, Cook R and Wilder D. Providing Government Information on the Internet: Experiences with THOMAS. In Digital Libraries Conference DL, 1995, pp. 19-24
- Jin Xu and Croft WB. Query expansion using Local and global document analysis, In Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. ACM, 1996. pp. 4-11.
- Attar R, Aviezri S Fraenkel. "Local feedback in full text retrieval systems", Journal of the ACM (JACM) 24. 3, 1977, pp. 397-417.
- Abdelmgeid Amin Aly, J. Using a Query Expansion Technique to improve Document Retrieval, International Journal "Information Technologies and Knowledge" Vol. 2, 2008, 343-348.
- Patricio Galeas, Ralph Kretschmer and Bernd Freisleben C. 2009. Document Relevance Assessment via Term Distribution Analysis Using Fourier Series Expansion. In JCDL '09: Proceedings of the 2009 Joint International Conference on Digital Libraries, ACM, 2009. pp 277–284.
- Patricio Galeas and Bernd Freisleben. J. 2008. Word Distribution Analysis for Relevance Ranking and Query Expansion. Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 2008, Volume 4919/2008, 500-511.
- Markus Holi, Eero Hyvonen and Petri Lindgren C. Integrating tf-idf weighting with Fuzzy view based search. In Proceedings of the ECAI Workshop on Text Based Information Retrieval, 2006.
- Sparck Jones K. , Automatic Keyword Classification for Information Retrieval. Butterworth, London, 1971.
- Peat Helen J, Peter Willet, The limitations of term co occurrence data for query expansion in document retrieval systems, JASIS 42. 5,1991,378-383.
- Luhn HP, The automatic creation of literature abstracts, IBM Journal of Research and Development, 2, 1958. 159-165.
- Rekha Vaidyanathan, Sujoy Das, Namita Srivastava. Query Expansion based on Equi-Width and Equi-frequency Partition. In Multilingual Information Access in South Asian Languages, Lecture Notes in Computer Science Volume 7536, 2013, pp 13-22
- Stephen Robertson, Understanding inverse document frequency: on theoretical arguments for IDF. , Journal of Docuemntation 60. 5,2004. 503-520.
- Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, Christina Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of the OSIR Workshop, 2006. pp. 18-25

Index Terms

Computer Science

Information Sciences
Keywords
Pseudo Relevance Feedback  Query Expansion  tf-idf  equi-frequency partition