General Formal Framework for Information Retrieval and its Applications

PhD Thesis

Tamás Kiezer

Supervisor: Dr. Sándor Dominich[†]
(1954 - 2008)

University of Pannonia

Faculty of Information Technology

Doctoral School of Information Science and Technology

2010

1 Introduction

With the advent of the Internet and World Wide Web (Web), Information Retrieval (IR) gained tremendous practical impact and theoretical importance. A number of retrieval methods have been elaborated since the inception, about half a century ago, which have been continuously evolving nowadays as well.

One of the classical methods is the Vector Space Model (VSM). It has been known for two decades that the VSM does not follow logically from the mathematical concepts on which it has been claimed to rest, but no proper solution has emerged so far. In this thesis, a general, discrepancy-free formal framework for IR is given and it is shown that using the concepts of this framework the Generalised Vector Space retrieval Model (GVSM), the Latent Semantic Indexing retrieval model (LSI) and the classical vector space retrieval model gain a correct formal mathematical formulation and background that is consistent with practice.

Based on this general framework the Entropy- and Probability-based retrieval methods are formulated consistently. Suited especially for the World Wide Web, the Combined Importance-based method is also derived from this

framework. A search engine called WebCIR is introduced, which implements this method.

Experimental evaluation results of the given methods are also reported. In vitro measurement of the Entropy- and Probability-based methods showed that, using these methods, improvement levels between 5 and 19 percent can be reached in comparison with the VSM and LSI methods. In vivo evaluation of the WebCIR search engine was also carried out. The results, which were compared to commercial search engines including Yahoo!, Altavista, and MSN, suggest that WebCIR is a very competitive retrieval and ranking technology.

2 Theses

The main contributions and the proposed theses can be summarized as follows:

1. *General formal framework for information retrieval*

Taking the definitions given for Information Retrieval (IR), they do not give different interpretations for IR, rather they all define IR the same way. In my dissertation a general formal framework for IR has been given.

(a) The concept of retrieval has been defined based on the mathematical measure theory. Then, documents (and queries) were particularised using fuzzy set theory [Chapter 3.2]. As a result, the retrieval function was conceived as the cardinality of the intersection of two fuzzy sets [Lemma 4.1].

(b) It has been shown that using the concepts of this general framework the generalised vector space retrieval model, the latent semantic indexing retrieval model and the classical vector space retrieval model gain a correct formal mathematical formulation and background that is consistent with practice [Chapter 4].

2. Entropy- and probability-based retrieval methods

The measure theoretic view (proposed in (1.a)) makes it possible to consistently formulate new retrieval methods. By taking fuzzy entropy and fuzzy probability as measures, new retrieval methods have been given, which are both consistent with their mathematical background.

(a) Entropy-based retrieval method has been given by taking fuzzy entropy as measure in the retrieval function [Chapter 5.1].

(b) Probability-based retrieval method has been given by taking fuzzy probability as measure in the retrieval function [Chapter 5.2].

(c) Effectiveness of the methods has been measured; experimental results using standard test collections have been reported. The experiments showed that enhancements from 5% to 19% can be obtained in average (over VSM and LSI), which indicates that the approach introduced in (1.a) offers a good basis for proposing new and better retrieval methods [Chapter 5.3].

3. Combined importance-based method for the retrieval and ranking of Web pages

Owe to the special properties of the World Wide Web, modern Web search engines typically use a mixture of retrieval methods partly based on classical methods and partly on properties of the Web graph.

(a) Using the concepts introduced in (1.a) and (2.b) a new method has been proposed for the retrieval and ranking of Web pages based on content importance, link importance, and topical similarity. The method is implemented in a search engine called WebCIR [Chapters 6.4 and 7].

(b) Four measurement methods have been used, involving human assessors as well, to evaluate the effectiveness of WebCIR, which was compared to the effectiveness of Altavista, Yahoo!, and MSN. The results show that the Combined importance-based method is a very competitive Web page retrieval and ranking method [Chapter 7.7].

3 Publications

Publications directly related to the thesis

[P1] DOMINICH, S., **KIEZER, T.**, ERDÉLYI, M. (2008). WebCIR: Web ranking and search engine using combined method. *Studies on information and knowledge processes 13.* Infota, pp.: 53-74. **[thesis 2, 3]**

[P2] DOMINICH, S., **KIEZER, T.** (2007). A Measure Theoretic Approach to Information Retrieval. *Journal of the American Society for Information Science and Technology*. John Wiley & Sons, Vol. 58, no 8, pp.: 1108-1122, ISSN 1532-2882, IF=1.773. **[thesis 1, 2]**

Other publications relevant to the thesis

[P3] DOMINICH, S., GÓTH, J., **KIEZER, T.** (2006). Web-based Neuroradiological Information Retrieval System using three methods to satisfy different user's aspect. *Computerized Medical Imaging and Graphics*, ISSN 0895-6111, pp: 263-272, IF=1.090.

[P4] DOMINICH, S., **KIEZER, T.** (2005). Hatványtörvény, „kis világ" és magyar nyelv. *Alkalmazott Nyelvtudomány*, pp: 5-25, ISSN 1587-1061.

[P5] DOMINICH, S., GÓTH, J., **KIEZER, T.** (2005). NeuRadIR: A Web-Based NeuroRadiological Information Retrieval System. *ERCIM News*, vol. 61., pp:52-53, ISSN 0926-4981.

[P6] DOMINICH, S., GÓTH, J., M. HORVÁTH, **KIEZER, T.** (2005). 'Beauty' of the World Wide Web – Cause, Goal, or Principle. *Lecture Notes in Computer Science*, Springer Verlag, Volume 3408/2005, pp:67-80, ISSN 0302-9743, IF=0.515.

[P7] DOMINICH, S., GÓTH, J., **KIEZER, T.**, SZLÁVIK, Z. (2004). Entropy-based interpretation of Retrieval Status Value-based Retrieval, and its application to the computation of term and query discrimination value. *Journal of the American Society for Information Science and Technology*. John Wiley & Sons, Vol. 55, no 7, pp: 613-627, ISSN 1532-2882, IF=1.773.

Citations

[C1] Bujdosó, I. (2006) Rangado – vortstatistika ekzamenado de la plurlingva teksto de la konstitucipropono de Europa Unio. *Proceedings Internacia Kongresa Universitato Florenco*, Italio, 29 julio – 5 augusto, pp: 134-143

[C2] Ianeva, T., Boldareva, L., Westerweld, T., Cornacchia, R., Hiemstra, D., and de Vries, A.P. (2004). Probabilistic approaches to video retrieval. *Proceedings of TRECVID International Conference*, National Institute of Standards, NIST, USA, pp: 1-10

[C3] Lafouge, T., Prime-Claverie, C. (2005). Production and use of information. Characterization of informetric distri-butions using effort function and density function. Exponential informetric process. *Information Processing and Management*, vol. 41, pp: 1387-1394, Elsevier, ISSN 0306-4573, IF=1,295

[C4] Janssens, F., Leta, J., Glanzel, W., Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management.* Elsevier, vol 42, no 2, pp: 1614-1642. ISSN 0306-4573, IF=1,215

[C5] Bordogna, G., Pagani, M., Pasi, G. (2006). A dynamic hierarchical fuzzy clustering algorithm for information filtering. *Studies in Fuzziness and Soft Computing*, Springer, vol. 197, pp: 3-23, ISSN 1434-9922.