Ghaleb Al-Gaphari
Linguistic Data Consortium LDC at the University of Pennsylvania in Philadelphia, USA.
drghalebh@yahoo.com

The Right Truncated Index-Based Web Search Engine

Abstract:

In this paper we investigate the result of applying the Right Truncated Index-Based Web Search Engine algorithm to determine its efficiency in terms of storing and retrieving Arabic documents.

The Right Truncated Index-Based Web Search Engine algorithm reads a set of Arabic documents and accepts a query, and then it processes both documents and the query. Hence it predicts the most relevant documents to the inserted query.

The Algorithm includes a morphological component (stemming algorithm) as well as a mathematical component (vector space model). The first component of the algorithm maps different inflected terms in an Arabic query into their original terms, thus minimizing the amount of required storage for the indexing system. At the same time, it relatively maximizes the probability of retrieving user favorable documents. The second component of the algorithm uses term frequency and inverse document frequency (TF-IDF) to determine the relative importance of each document based on the terms of the query.

The TF-IDF (term weighting schema) computes the weight for each term in an Arabic document by multiplying the inverse document frequency array by the term frequency array. It then computes the cosine similarity between the query vector and each document vector in the collection. The greater the cosine similarity between the query terms and the document terms, the greater the relevancy the document has to the query. In other words, the greater the cosine similarity between the query terms and the document containing those terms, the greater the probability that the document will be of user interest, thus improving the query retrieval.

References:

[1] Kareem Darwish ,"Building a shallow Arabic Morphological Analyzer in One Day ",in ACL02 workshop on Computational Approaches to Semitic Languages,Philadelphia,PA, Association for Computational Linguistics,2003.

[2] Leah S. Larkey,Lisa Ballesteros, and Margaret E. Connell,"Improving Stemming for Arabic Information Retrieval:Light Stemming and Co-occurrence Analysis ",*SIGR'02 Augest 11-15 , 2002,Tampere Finald.*

[3] Leah S. Larkey and Margaret E. Connell," Arabic Information Retrieval at UMass in TREK-10 "

TRECK 2001.Gaithersburg:NIST,2001.

[4] Tim Buckwalter" Buckwalter Arabic morphological analiser version 2.0 ",2004.

[5] M, Aljlayl, O. Frieder, and D. Grossman," On Bidirectional English –Arabic Search",JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,53(13):1139-1151,2002.

[6] Jinxi Xu,Alexander Fraser ,and Ralpha Weischedel " Empirical Studies in Strategies for Arabic Retrieval " SING'02,Augest 11-15,2002,Tampere,Filand.

[7] John Broglio,James P. Callan, and W. Bruece Croft, "Inquery System Overview",TREC-9 Gaithersburg,Maryland,2000.

[8]Haider Moukdad, "Lost In Cyberspace:How Do Search Engine Handle Arabic Queries ?", Available at http://www.morfix.com/arabicSearch/arabic/help.html, ,2004 .

[9] Nizar Habash , Owen Rambow, and George Kiraz, "Morphological Analysis and Generation for Arabic Dialects",Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages

17-24,AnnArbor,June 2005.

[10] Nizar Habash and Owen Rambow, " Arabic Tokenization,Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop",Proceedings of the 43rd Annual Meeting of the ACL,pages 573-580,Ann Arbor,June 2005.©2005 Association for Computational Linguistics.

[11] Ricardo Baeza-Yates and Carlos Castillo, " Web Search " Proceedings of the third Workshop on Web Graphs (WAW) ,volume 3243 of Lecture Notes in Computer Science ,pages 156-167,Rome,Italy.Springer 2005.

[12] J. Michael Schultz,Mark Y. Liberman, " Towards a Universal Dictionary for Multi-Language Information Retrieval Applications " Available at http://www.ldc.upenn.edu,May,22,2000

[13] Julie Weeds and David Weir, " Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity " ,©2006 Association for Computational Linguistics.

[14] Alon Y. Levy, Anand Rajaraman,and Joann J. Ordille" Querying Heterogeneous Information Sources Using Source Descriptions "Proceedings of the 22nd VLDB Conference Mumbai (Bombay) ,India,1996.

[15] Mehmet Altinel and Michael J. Franklin , ” Efficient Filtering of XML Documents for Selective Dissemination of Information “,Proceedings of the 26th VLDB Conference Cairo, Egypt ,2000.

[16] Douglas W. Oard, ” The TRECK-2002 Arabic/English CLIR Track“,SIGIR’02,Augest 11-15, 2002,Tampere,Filand.

[17] Marc Najork,and Allan Heydon, ” High-Performance Web Crawling“,©2001,Kluwer Academic Publishers,Inc.