The Next Generation of Information Retrieval Applications

João Ferreira
ISEL
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, IST
alberto.silva@acm.org

José Delgado
Instituto Superior Técnico
Jose.Delgado@tagus.ist.utl.pt

ABSTRACT

We describe an Information Retrieval (IR) framework to cover the most relevant IR models based on statistics and link properties of the documents. We also propose to use the same framework for IR, filtering and classification process.

KEYWORDS

Framework, Information Retrieval, Standard, algorithms, IR models.

1. Introduction

Between 2000-2005, the foundations and concept of the retrieval framework that can unify all statistical / mathematical IR models have been developed under the name of WebSearchTester. We do not intend to propose one more IR application, but rather to build one that can integrate all IR models based on a flexible index, layer and modular structure. The main purposes of the WebSearchTester framework are:

- To create a common platform that support the following IR models: Boolean, Vectorial, Document generation (classical probabilistic, implemented Okapi measure), Query generation (Languages model), Logistic Regression (LR), Inference network, Concept space model, Probabilistic distribution, Link analyses, KL divergence, Markov chain, Fusion of results from different models and also Classification. Details of this implementation can be found in [1]; (1) to make available better and comparable evaluation of different models based on a common infrastructure; (2) to promote research in a collaborative environment; (3) to propose a common framework for IR, filtering and classification applications; (4) to support personalization analysis that can be used to rank IR results in terms of users' preferences; (5) to support automatic community identification, to enable the mass consumers and to progress towards to diffusion systems.

2. WebSearchTester

The system contains a number of modules, which can be organized around 3 layers, developed in a distributed "plug-in" architecture. The first layer is responsible for creating information representatives and user interaction; the second layer is the layer of IR models; the third layer is application oriented. Appropriate communication space allows communication between different layers.

First layer: This layer has the objective to feed the second layer with valuable data by creating and storing data representative of information and user information needs. It contains the following modules:

- The indexing module, created in a flexible way, is divided in the following sub-modules; (1) parser, user can choose different fields to index (e.g., header terms, full document, URL, phrase (with or without dictionary) and more); (2) Stop-list (several stop-word list available in different languages); (3) stemming (we implement snowball <snowball.tartarus.org>); (4) weighting (several weight schemes available, see [1]),

(5) storage perform through openfts <openfts.sourceforge.net>in a postgresql database and text files. Index contains terms representatives of documents, URL, statistic proprieties of documents, terms and collections.

- The feedback module expands the initial query in a pseudo-feedback retrieval process based on the adaptive linear model using the top ten positive and top two negative weighted terms, from the top three ranked documents of the initial retrieval results.
- The user interface module is responsible to get and to present information from the user. An interface to classified systems is also available.
- The crawler module, picks information over the internet, when needed (no controlled collection used) and was integrated Larbin , <larbin.sourceforge.net>.
- The translator module (allows multi-language search), integrated Wordtrans <wordtrans.sourceforge.net>, providing translations in six different languages: German, French, Italian, English, Spanish and Portuguese.

Second layer: In this layer we propose an open implementation (figure 1) of the most relevant IR models based on statistics and link proprieties of documents. This layer provides relevant metrics regarding documents and user queries through different statistics approaches and based on flexible index scheme. Index information contains pure statistics information (e.g., term frequency, number of terms in a collection, document length and others). All weights and estimation functions are performed on this layer because they are depend on the IR model used.

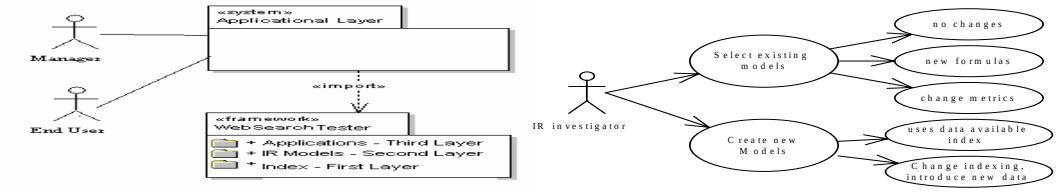

Figure 1: Framework main concept (left) and use case of IR investigator of WebSearchTester.

Third layer: This layer is oriented to the specific application, and includes the following modules: (1) profile builder, handles information regarding user stable information needs, periodicity, mail and also a negative profile related with not relevant information; (2) community identification, based on the fusion of the following methods: clustering profiles, documents and also link analysis of community identified relevant documents [2]; (3) fusion, we implemented two new approaches based on the most common fusion formulas with small changes, the Similarity Merge (SM) and Weighted Sum (WS) [3]. We propose also a mixed formula to use both approaches, see [1]; (4) classification, we implement KNN (k nearest neighbors), SVM (support vector machines), TM (term match), LLSF (linear least square fit) and NB (naïve bayes). Details can be found at [1]; (5) classified system (CS), several CS implemented: ACM, AMS, Web Directory (Yahoo!).

Applications: IR systems have been tested under a controlled environment, using TREC WT10g collection, Topics 451 to 550 and relevance judgment. Results are divided in: (1) IR models; (2) internal systems parameters; (3) fusion of internal systems; (4) fusion of different IR models. We performed more than 3000 systems tests from 2003 to 2004. Filtering systems performed through a personalized TV and Newspaper [4].Classification systems implemented through Mydocument, a system to classify enterprise information [5]. User profiles were used as pseudo feedback to tune results from different IR models.

REFERENCES

[1] www.deetc.isel.ipl.pt/matematica/jf/ir.hmtl

[2] J.Ferreira, A.Silva and J. Delgado. Fusion methods to find Web Communities, In Proceedings of the Web based Communities 2005, 23-25 of February 2005, Alvarge.

[3] Lee J. H. Analyses of multiple evidence combination. Proceedings of the ACM SIGIR, 1997, 267-276.

[4] J.Ferreira, A.Silva and J. Delgado. Personalized filtering Systems Based on the Combination of Different Methods. In Proceedings of Applied Computing 2005, 22-25 of February 2005, Alvarge.

[5] J.Ferreira. Retrieval information over internet, Phd Thesis, Setember 2004 (in Portuguese).