PIFA: Designing a Personalized Information Filtering Algorithm for Knowledge Management Systems

Olusegun Folorunso, Rebecca Olufunke Vincent, Oni Oluwaseyi Akanji
Department of Computer Science,
University of Agriculture Abeokuta, Ogun State, Nigeria

Adewale Opeoluwa Ogunde
Department of Mathematical Sciences,
Redeemer's University (RUN), Redemption City, Mowe,
Ogun State, Nigeria

ABSTRACT. A study on the concept of "personalized information filtering" system was carried out. Natural Language Processing (NLP) was used to tag the words, and metrics such as TF-IDF was used to weigh each term in the document. Relevance feedback was used to get users' judgments. The approach promises to push relevant information directly to the user in a timely and efficient manner.
Keywords: Knowledge, Knowledge Management, Information filtering, Natural Language Processing, Vector Space Model

Introduction

As tremendous amount of information is created and delivered over electronic media, effective management of the potentially infinite flow of information is getting increasingly difficult for individuals to control. Similarly, just as more and more users are getting online, it is getting increasingly difficult for people even experts to find information unless one knows exactly where to get it from and how to get it. It is however more likely that an expert involved in a knowledge intensive task would recognize the usefulness of the information he or she receives.

Knowledge Management is a technique used by organizations and communities to improve how business is conducted by leveraging data and

information that are gathered, organized, managed, and shared. Knowledge management could also be referred to as the management discipline that is concerned with creating, preserving, and applying the knowledge that is available within an organization. Typically, this is done using techniques such as training, process modeling, experience databases, and networking of experts.

Information filtering is an approach for sourcing for useful materials online, filtering irrelevant ones and bringing the relevant ones to a user's desktop. It is a push approach that pushes knowledge to the user (connecting knowledge to people). It is one of the three disciplines that try to cope with the problem of information overload and information mismatch due to the large amount of digital information available on the World Wide Web. Others are Information Retrieval and Text Categorization.

Tools to regulate the flow are urgently needed to prevent computer users from being drowned by the flood of incoming information [MIT09]. One of such tools is "Information Filtering". According to [Z+07], "information seekers are unable to specify their information needs precisely and accurately due to the lack of training or unfamiliar with the collection makeup and retrieval environments.

Queries submitted to search engines by Web users are generally very short and contains only several keywords. These short queries are not able to reflect the underlying intentions of web users. This often leads to information overload and information mismatch problems. If large amount of returned hits must be gone through manually, Web users are likely frustrated and they may not get any useful information. Therefore, it is very desirable to develop efficient personalized information gathering systems to meet what users want, which is the major focus of this paper.

1 Literature Review

Many researches have been carried out by several people to deal with web information overload and mismatch problems. These are done in order to filter online information and deliver the relevant ones to the user. These researches include topics in knowledge management, information retrieval, information filtering, etc. to mention but a few. This review talks briefly on the purpose of research in an academic environment, it also discusses the discipline of knowledge management, followed by the concept of "push technology" which automatically provides readers with information via the internet without them specifically requesting for it. It then goes on to discuss the on-going research in the technique of information filtering.

1.1 Information Filtering

Information filtering could be defined as a push approach that pushes information to the user (connecting knowledge to people). It is one of the three disciplines (others are Information Retrieval and Text Categorization) that try to cope with the problem of information overload and information mismatch due to the large amount of digital information available on the World Wide Web. As a result of their common goal, these three disciplines have some characteristics in common but a lot of differences as well [Nan01].

An information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation to a human user [Wik09]. Information filtering systems can help users by eliminating the irrelevant information and by bringing the relevant information to the user's attention. Filters are mediators between the sources of information and their end-users. Information Filtering is based on the requirements that:

a) It should be able to assess relevance of information entities based on their content and not on some surrogate representation of this content;
b) It should be able to minimize the individual's interaction with the system; and
c) It should be able to adjust to changes in the needs and interests of the user.

The focus of Traditional Information Retrieval research is the development of algorithms and models for the retrieval of textual information from document repositories. Text Categorization is concerned with the problem of automatically assigning a class label or subject descriptor to texts that belong to the same class and thus their subsequent retrieval. Information Filtering is also an information access activity that deals with the filtering of a dynamic stream of incoming textual information, according to evolving user interests or needs. Information filtering systems can help users by eliminating the irrelevant information and by bringing the relevant information to the user's attention. Filters are mediators between the sources of information and their end-users [Pal00].

1.2 Related Works on Information Filtering

The discipline of personalized information filtering, that has only recently become fashionable with the emergence of intelligent information agents and personal assistants, is now entering the domain of knowledge management [GAD98]. The *Knowledge Pump* system for example, uses

community-centred collaborative "trust" between users within a given category of documents [GAD98]. According to this approach, a document is presented to a user if it was highly rated by another individual whose judgements the user trusts.

A more interesting approach is followed by the *Knowledge Sharing Environment (KSE)* system. KSE is a system of information agents for organizing, summarizing, and sharing knowledge from a number of sources, including WWW, an organization's intranet or from other users [DSW98]. Each user has his information agent, which maintains a user profile that represents the user's information needs and interests. The user profile comprise a number of user specified phrases or terms and is refined according to its usage as it will be explained below. KSE agents provide a number of knowledge management services.

The role that information filtering technology can play in the development of knowledge management systems was investigated by [Nan01]. He stated that knowledge-based systems suffer from a number of disadvantages in the way personalization of information delivery is supported, and that technology and human-oriented considerations indicate that to support the delivery of relevant enough information, a KM system has to be flexible enough to changes in the individual's information needs as these are implied by his task at hand.

For the realization of an Information Filtering system for the Web, [MMS98] presented a system capable of selecting HTML/text documents, collected from the Web, according to the interests and characteristics of the user. The system was based on a hybrid architecture, where an artificial neural network is integrated into a case-based reasoner. The system, based on a user modelling component, is designed for building and maintaining long term models of individual Internet users, and acts as an intelligent interface for the Web search engines. It selects the documents according to the *context* interests (and non-interests) of the user, as desumed by the system through the interaction, making use of a User Modelling ad-hoc subsystem, particularly conceived for Internet users. The vector space model was used as the basics, made up of sets (called clusters) of pairs (*term*,), with *term* a word and *context* an object whose purpose is to disambiguate possibly ambiguous terms using words before and after that word. An evaluation of their system was first conducted through real-time access to the World Wide Web.

In dealing with web information and mismatch problems, [Z+07] presented a web information gathering method which integrates the search intent based filtering and pattern based data mining technology together to alleviate Web information overload and mismatch problems. The design of

the web information system included two phases: the ontology-based user profiling, and pattern discovery. The information filter, based on user search intents, was used to quickly filter out the likely irrelevant data, and then a data processing was carried out on the "cleaned" – reduced data by improving upon the orthodox data mining method, pattern taxonomy (PTM), with their new method called Filtering-based Web Information Gathering (FWIG), that rationalizes the data relevance.

2 Methodology

The profiles used by the filtering system consist of terms which are matched with the contents of the documents. Research in Artificial Intelligence, basically Genetic Algorithm and Artificial Evolution motivates the learning mechanism. The learning mechanism used by intelligent agents is relevance feedback [JTW07] and genetic algorithm [Z+07]. According to [MIT09], Information filtering is effectively a dynamically changing search problem. In this project, relevance feedback would be used as the learning mechanism.

2.1 The Vector Space Model

The Information Filtering agent is modeled as a population of profiles (profile individuals) [Z+07]. The representation used for profiles and documents is based on the vector space representation, commonly used in information retrieval literature. In the vector space representation, documents and queries are both represented as vectors in some hyper-space. A distance metric which measures the proximity of vectors to each other is defined over the space. When a query is received, it is translated into its vector representation and document vectors in the proximity of the query vector are retrieved in response to the search. The advantage of using a common vector space for both documents and queries is that a document can also be used as a query itself i.e. one can find documents that are similar to a given document. Once the document-query is translated to a vector, the same distance metric can be used as for other queries. This property of vector spaces is quite useful for the current application, since users can provide samples of interesting articles as an alternative to constructing intelligent queries.

Profiles are analogous to queries in information retrieval [MIT09]. Usually, a profile searches part of the database looking for articles that are similar to it. It searches for documents that match itself and recommends

them to the user. The user can provide feedback for the documents recommended. Profiles and documents' representations are described below:

Documents

According to [BDS94], a "term" is used for text identification. Since the terms are not all equally important for content representation, importance factors (or weights) are assigned to the terms in proportion to their presumed importance for text content identification. A text is then representable as a vector of terms where *w* represents the weight of term in text. The documents can contain other information such as the author of the document, the source of information, etc. It is therefore necessary to generalize the term-vector representation mentioned above. The generalized representation used is as follows:

A document consists of many fields. Each field is assigned terms to be used for identifying purposes. Since the terms are not all equally important, they are assigned weights. A field is thus represented as a term-vector.

w_{ij} is the weight of term in field. The subscript i can be "a"(uthor), "k"(eyword), "l"(ocation) and so on. The superscript indicates that F_i^d is a document field, (as opposed to a profile field described below).

Since a document consists of many fields, it is represented as a set of field-vectors. Formally,

$$D = F_i^d \qquad \text{(Eqn 1)}$$

where D is a document and F_i^d is a field in the document.

Profiles

A profile consists of a number of fields, like author, location, keyword, etc. Each field is a vector of terms, each of which is weighted in proportion to its importance for identification purposes. The profile stands for some user interest and it is a set of fields with field-weights,

$$p = \{(F_i^p, W(F_i^p))\} \qquad \text{(Eqn 2)}$$

where $W(Fp_i)$ gives the weight of the field F_i^p in profile P. The superscript p indicates that F_i^p is a profile-field and not a document field. Each profile field F_i^p is represented identically to a document field:

$$F_i^p = < w_{ij}^p > \qquad \text{(Eqn 3)}$$

Field Extraction

The fields of the document representation must be extracted from the document itself. All document fields except the keyword field are directly extracted from the header lines of the article. The keyword field is generated

from the text of the article. The term-vector for the keyword field is obtained through a full text analysis of the documents.

A well known term weighting method for the vector-space model in information retrieval literature is adapted for this information filtering where the weight of a term depends on its frequency of occurrence in the text and the number of documents it appears in. The weight of a keyword-term is the product of its term frequency (*tf*), which is the occurrence frequency of the term in the text (and it is normally reflective of the term importance), and its inverse document frequency (*idf*), which is a factor which enhances the terms which appear in fewer documents, while downgrading the terms occurring in many documents [Nan01], [MMS98]. The weight w_{ik} of the term t_k is given as

$$w_{ik} = tf_{ik} \times idf_k \quad \text{(Eqn 4)}$$

Where tf_{ik} is the number of occurences of term t_k in document i, and idf_k is the inverse document frequency of the term in the collection of documents [BLP00]. A commonly used measure for the inverse document frequency is

$$idf_k = 1 + \log\left(\frac{N}{df_k + 1}\right) \quad \text{(Eqn 5)}$$

Where N is the total number of documents in the collection in which term t_k appears once or more; df_k defines the number of times that term t_k appears in the collection. This boosts the influence of rare terms [BLP00]. The collection of documents is the context within which the inverse document frequencies are evaluated.

Learning from feedback

There are two ways in which the user can communicate with the agent by providing feedback about interesting articles. The first way is to provide positive or negative feedback for the articles retrieved by the agent (or one of its profiles). Secondly, the user can provide examples of articles that the agent did not retrieve, which is an example of programming by demonstration. In this project, the user will be making use of the first approach.

2.2 The Natural language Processing (NLP) library

Tagging of words was done through a lexicon tool derived from the knowledge within the NLP library. OpenNLP is both the name of a group of open source projects related to natural language processing (NLP), and the name of a library of NLP tools written in Java by Jason Baldridge, Tom

Morton, and Gann Bierner. The tool works by getting the words in a document and categorizing them into their various parts of speech representation, such as noun, adjective, verb, etc. Part-of-speech tagging is the act of assigning a part of speech (sometimes abbreviated POS) to each word in a sentence [OB10].

Having obtained an array of tokens from the tokenization process, we can feed that array to the part-of-speech tagger. The POS tags are returned in an array of the same length as the tokens array, where the tag at each index of the array matches the token found at the same index in the tokens array. The POS tags consist of coded abbreviations conforming to the scheme of the Penn Treebank, the linguistic corpus developed by the University of Pennsylvania. It uses tags such as shown in figure 2.1 below to tag the words, and it is a useful tool in analyzing the document to be used as a query.

CC Coordinating conjunction RP Particle
CD Cardinal number SYM Symbol
DT Determiner TO to
EX Existential there UH Interjection
FW Foreign word VB Verb, base form
IN Preposition/subordinate VBD Verb, past tense
conjunction
JJ Adjective VBG Verb, gerund/present participle
JJR Adjective, comparative VBN Verb, past participle
JJS Adjective, superlative VBP Verb, non-3rd ps. sing. present
LS List item marker VBZ Verb, 3rd ps. sing. present
MD Modal WDT wh-determiner
NN Noun, singular or mass WP wh-pronoun
NNP Proper noun, singular WP$ Possessive wh-pronoun
NNPS Proper noun, plural WRB wh-adverb
NNS Noun, plural `` Left open double quote
PDT Predeterminer , Comma
POS Possessive ending " Right close double quote
PRP Personal pronoun . Sentence-final punctuation
PRP$ Possessive pronoun : Colon, semi-colon
RB Adverb $ Dollar sign
RBR Adverb, comparative # Pound sign
RBS Adverb, superlative -LRB- Left parenthesis *
-RRB- Right parenthesis *

Figure 2.1: POS tags Listing by Pen Treebank
Source: [OB10]

OpenNLP is both the name of a group of open source projects related to natural language processing (NLP), and the name of a library of NLP tools written in Java by Jason Baldridge, [OB10]. It provides a number of natural language processing tools based on maximum entropy models. The tools used in the C# programming language are: a sentence splitter, a tokenizer, a part-of-

speech tagger, a chunker (used to "find non-recursive syntactic annotations such as noun phrase chunks"), a parser, and a name finder.

2.3 The Personalized Filtering Model

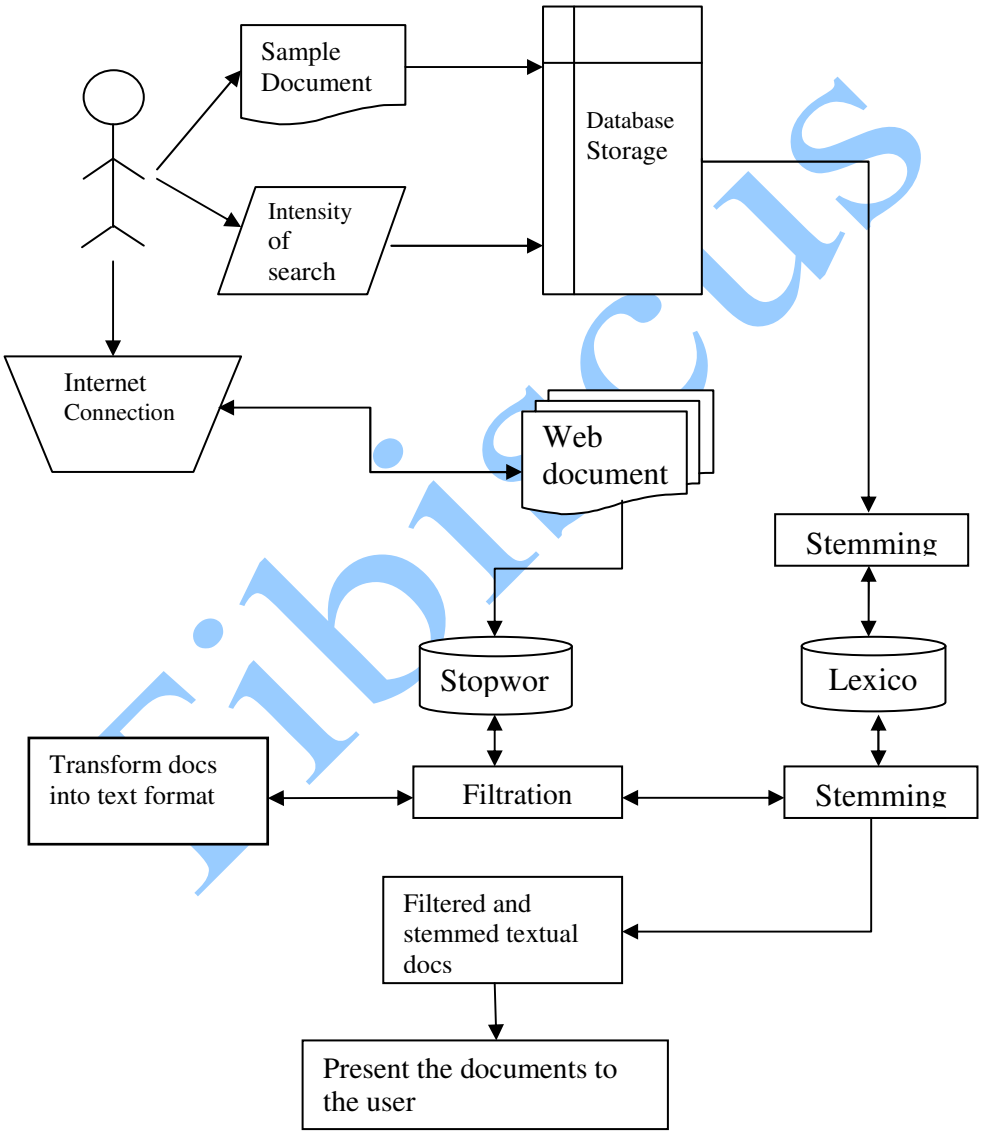
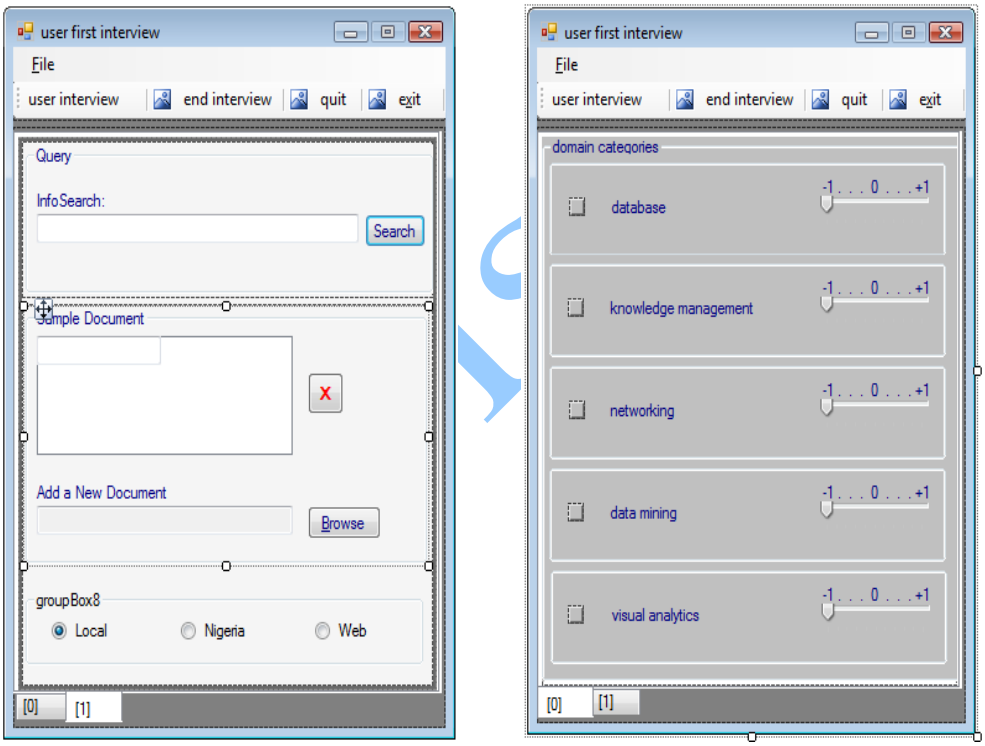


Figure 2.2. The Personalized Information Filtering Model

3 System Implementation and Performance Evaluation

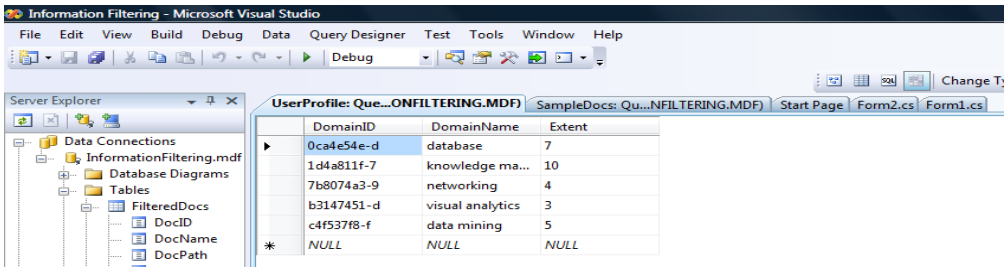
The proposed system described in section 2 was implemented in C-sharp (C#) language using Microsoft Visual Studio 2008 in Microsoft Vista operating system. Figures 3.1 (a and b) represent the forms needed for the request for few pieces of personal information from the user based on the domain categories submitted by the intended user to the designer, prior to the release of the system, for example, the extent to which he wants information on the categories, a sample document to serve as an alternative to constructing intelligent queries, the search region (whether local, web, etc, (and other areas of interest, if necessary).



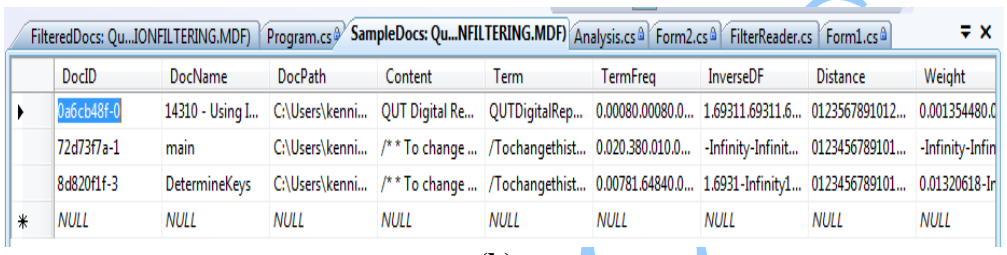
(a) **(b)**

Figure 3.1. User's first interview

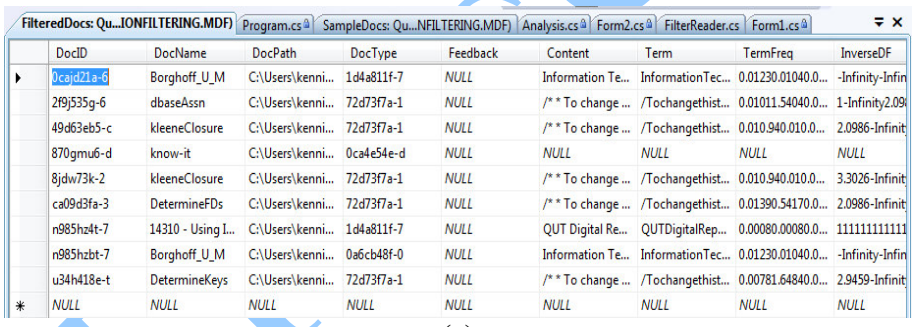
This information, collectively taken as user profile, is then entered into a "server-side" database (generally, a "server side" database is a database that is kept on a push content provider's computers or "servers (figure 3.2).



(a)



(b)



(c)

Figure 3.2. Information Filtering Database

This database is then used to determine what information the user might be interested in and when that information should be sent to the user. To know the search intent of the user, the user has to first fill in a form – user profile, stating the extent to which he wants information on the interested domains of knowledge. He can also supply sample documents as an alternative to constructing intelligent queries (Figure 3.3).

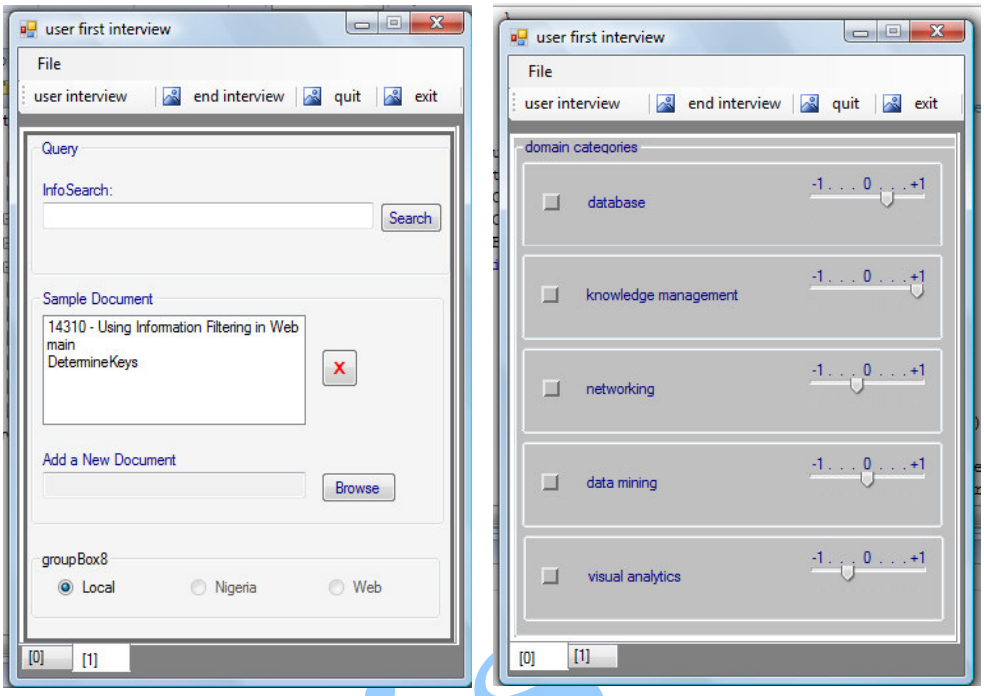


Figure 3.3. User Profile containing the user's information need

As the system is not assumed to be omniscient or omnipotent, its role is to present information it "believes" would be of best interest to the user, but it is the user that decides whether a document is really of great value or not. Hence, relevance feedback is used to get user's judgement of the presented information. The feedback for a particular document is received at the upper right corner of figure 3.4. The NLP tool is used to tag each word in the presented document, and determines the keywords that best describes the document.

Conclusion and Future Works

This study discussed the technology behind the automatic distribution of selected data to user's computer at prescribed intervals without them specifically requesting for it, such as the e-mail. The concept of "personalized information filtering" system was examined, which is a "push" concept, to automatically filter out irrelevant materials and bring the relevant ones to the user's attention based on the user profile representing

the user's interests and needs, so as to better reflect on his decision making processes and thus acquire new knowledge; and thus providing added convenience for the knowledge seeker.

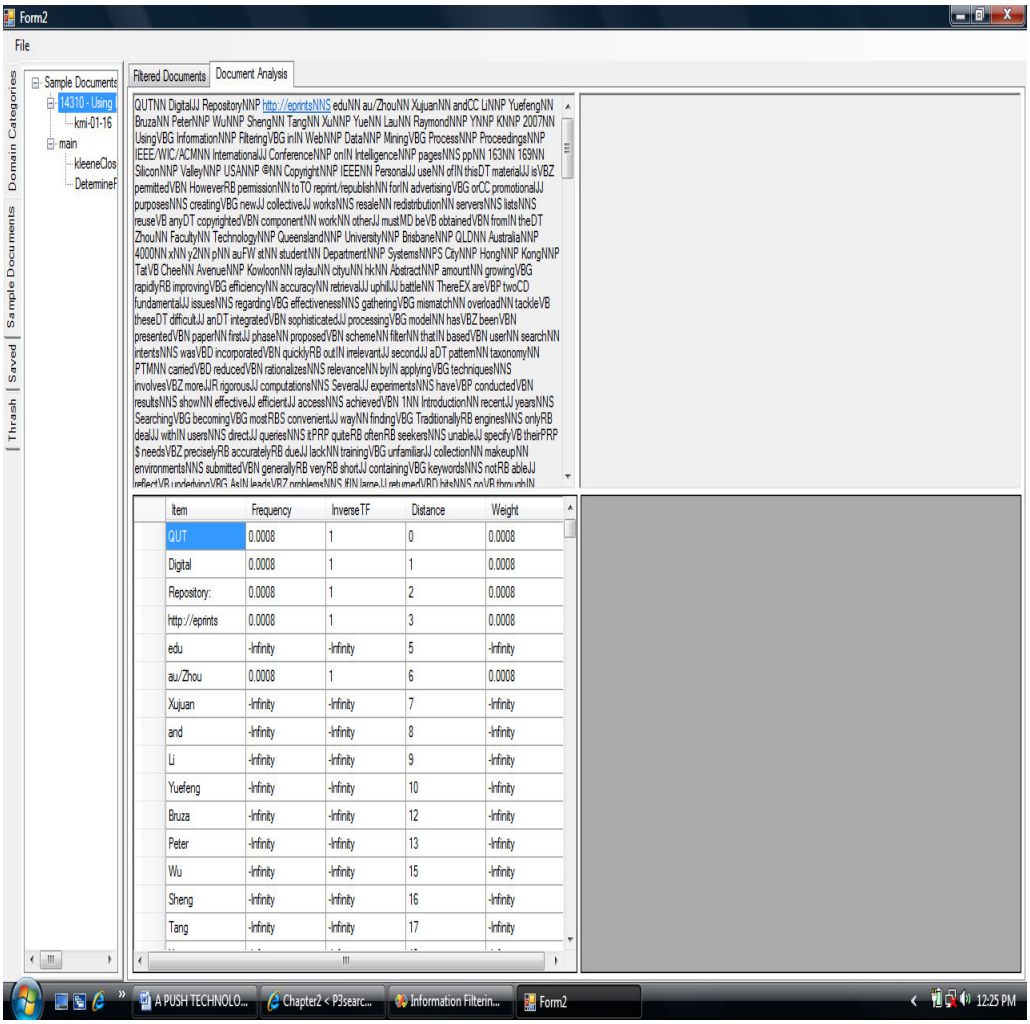


Figure 3.4. Document Analysis

The approach was to request for a few pieces of personal information from the user, for example, the extent to which he wants information on his stated categories, the kind of articles he is interested in, a sample document to serve as an alternative to constructing intelligent queries, and the search region (An information filtering system may require more as needed). A lexicon tool from Natural Language Processing (NLP) was used to tag the

words in a document breaking them down into their various parts of speech, and metrics such as TF-IDF was used to weigh each term in the document. Relevance feedback was used to get user's judgment of the presented information. The result of the work shows that instead of having information seekers to visit a corporate website only to have to hunt for relevant information, this technology holds the promise of pushing that relevant information directly to the user.

Finally, the system designed in this work dealt with the problem of information mismatch and information overload by filtering irrelevant materials and bringing the relevant ones to the user's attention. It provided added convenience for the knowledge seeker by minimizing user's interaction with the system. There is also the possibility of bypassing the web altogether. Future works intends to build standardized architecture for filtering systems upon which other systems can be built. Also, enabling the overall security of the user's system is another area to consider for future work.

References

[BLP00] **S. M. Bohte, W. B. Langdon, H. L. Poutre** - *On Current Technology for Information Filtering and User Profiling in Agent-Based Systems*, 2000.

[DSW98] **J. Davies, S. Stewart, R. Weeks** - *Knowledge Sharing over the World Wide Web*, WebNet '98, Florida,USA, November 1998.

[GAD98] **N. Glance, D. Arregui, M. Dardenne -** *Knowledge pump: Supporting the flow and use of knowledge*, in Information Technology for Knowledge Management (U. Borghoff and R. Pareschi, eds.), ch. 3, Springer-Verlag, 1998.

[JTW07] **I. Jung, D. Thapa, G. Wang** - *Intelligent Agent Based Graphic User Interface (GUI) for e-Physician*, World Academy of Science, Engineering and Technology 36, pages 194-197, 2007.

[MIT09] **MIT Media Lab** - *Autonomous Agents Group - agentmaster@media.mit.edu*

http://agentmaster@media.mit.edu assessed on November 17, 2009

[MMS98] **M. Marinilli, A. Micarelli, Sciarrone** - *A Case-Based Approach to Adaptive Information Filtering for the WWW*, 1998.

[Nan01] **N. Nanas** - *Information Filtering for Knowledge Management*, 2001.

[OB10] **T. Ollar, J. Bennett -** *Design concept for a new Visual Studio UI*,.Visual Studio 2010 Concept IDE, The Code Project 2010.

[Pal00] **J. Palme** - *Information Filtering*, in Proc. of the ITS 2000 conference.

[She94] **B. D. Sheth -** *A Learning Approach to Personalized Information Filtering,* M.Sc. thesis, Submitted to the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, 1994.

[Wik09] **Wikipedia, 2009 -** *Information filtering system*, accessed December 7, 2009

[Z+07] **X. Zhou, L. Yuefeng, P. Bruza, S. Wu, Y. Xu, R. Lau -** *Using Information Filtering in Web Data Mining Process*, Proc. IEEE/WIC/ACM International Conference on Web Intelligence, pages pp. 163-169, Silicon Valley, USA, 2007.

http://www.babylon.com assessed on November 8, 2009

http://www.LearntheNet.com assessed on November 7, 2009

http://www.wikipedia.org assessed on November 7, 2009

http://www.zdnet.com assessed on November 7, 2009