Adaptive information retrieval system based on fuzzy profiling

Alhabashneh, O. , Iqbal, R. , Doctor, F and Amin, S.

Original citation:
Alhabashneh, O. , Iqbal, R. , Doctor, F and Amin, S. (2015) 'Adaptive information retrieval system based on fuzzy profiling' in 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp: 1-8). IEEE. DOI: 10.1109/FUZZ-IEEE.2015.7338012

http://dx.doi.org/10.1109/FUZZ-IEEE.2015.7338012

IEEE

Adaptive Information Retrieval System Based on Fuzzy Profiling

Obada Alhabashneh, Rahat Iqbal, Faiyaz Doctor, Saad Amin
Computing Department
Coventry University
Coventry, UK
o.alhabashneh@coventry.ac.uk; r. iqbal@coventry.ac.uk; faiyaz.doctor@coventry.ac.uk ; s.amin@coventry.ac.uk

Abstract— **The importance of finding relevant information for business and decision making is imperative for both individuals as well as enterprises. In this paper, we present an approach for the development of a fuzzy information retrieval (IR) system. The approach provides a new mechanism for constructing and integrating three relevancy profiles: a task profile, a user profile and document profile, into a unified index through the use of relevance feedback and fuzzy rule based summarisation. Experiments were performed from which relevance feedback and user queries were captured from 35 users on 20 predefined simulated enterprise search tasks. The captured data set was used to develop the three types of profiles and train the fuzzy system. The system shows 86% performance accuracy in correctly classifying document relevance. The overall performance of the system was evaluated based on standard precision and recall which shows significant improvements in retrieving relevant documents based on user queries.**

I. INTRODUCTION

The amount of digital information available on the Internet and various Intranets often causes information-overload, significantly increasing the amount of time and cognitive resources needed to acquire relevant and accurate information. Current enterprise systems produce results based on specific keywords without taking into account user's context, such user location, browsing history, and previous interaction patterns. The research performed by the International Data Corporation on information workers showed that more than 26% of their search sessions failed to bring any relevant results [1]. It was estimated that the information workers spend approximately 9% of their overall time searching for information that did not produce any results. This leads to a decreased quality of products as well as decisions being based on inaccurate or out of date information [2].

In order to produce accurate search results corresponding to the needs of a user, it is necessary to develop an intelligent IR system. Such systems can be developed using relevance feedback based approaches [3] that are based on the knowledge of how relevant the particular piece of information (document) is to the user and how its content can be reused in order to find documents that are similar. Documents that are similar to the relevant content have a very high probability of being returned or retrieved. There are two techniques of relevance feedback: explicit and implicit [4]. In explicit feedback, users explicitly mark the documents as relevant or not whereas in implicit feedback, the relevance is estimated based on behavioural observation such as reading time, click count, etc.

User profiling can be developed using the above mentioned relevance feedback approaches. User profiling is one of the significant techniques in modern IR systems where such profiles contain user browsing history, tasks, preferences and interest [5].

The modern IR systems should be self-learning and adaptive by responding accurately and timely to user needs. These systems mainly use machine learning techniques to learn and adapt their models over time [6]. Fuzzy logic can be used to enhance the classification of user relevancy by handling the uncertainty and ambiguity in user data. Fuzzy sets provide an expressive method for user judgment modelling and fuzzy rules provide an interpretable method of representing the classification rules.

In this paper, we present an approach for the development of a fuzzy based IR system. The approach provides a new mechanism for constructing and integrating three relevancy profiles: a task profile, a user profile and document profile, into a unified index through the use of relevance feedback and fuzzy rule based summarisation [7].

We used the relevance feedback to develop a linear predictive model showing the association between the implicit and explicit feedback parameters. The model was used to predict the document relevancy from the implicit user feedback parameters. The predicted relevance values were used to identify the successful queries (which led to document visits) and train the fuzzy rule summarising model. The successful user queries were preprocessed and the query terms were extracted. TF-IDF (Term Frequency and Inverse Document Frequency) matrices were calculated for the terms and used by the fuzzy system to create profiles for task, user and document. For each of these profiles, each search term was then associated with its retrieved documents and the predicted relevancy (i.e. term weight) . This formed a rule base consisting of three inputs (term weights associated with the three profiles) and one output which was the predicted relevance level of the document. Then the fuzzy rule based summarisation was applied to extract the most representative fuzzy rules. These were used to build the unified relevancy index. A web-based user interface was developed to handle the user queries and display results based on the user query.

The rest of the paper is organised as follows: Section II presents a literature review of relevance feedback and fuzzy logic approaches for IR systems. Section III

describes the proposed method. Section IV discusses experiments and results. Finally conclusions and future directions are presented in Section V.

II. LITERATURE REVIEW

Towards the development of intelligent IR systems and user profiling, relevance feedback has been investigated by several researchers [3] [8] [9] [10]. Previous research has analysed user behaviour and found a significant relationship between the time spent on reading Usenet news and interest level. This was proved by comparing observational studies with explicit interest measures.

Current research shows that the combination of several relevance feedback parameters can produce better results [11], [12], [13], [14] and [15]. It was found that reading time, along with some other user behaviours can be a very reliable indicator of content relevancy. It was noticed that even though there is a positive correlation between mouse movement and amount of clicks, reading time was shown to be a reliable indicator of user interests [16]. In [15] multiple implicit parameters (dwell time, click-through, text selection and page review) were combined to investigate their impact on the document relevancy. The experiments showed that the retrieval performance was improved when more parameters were used.

The relationship between user behaviour during the dwell time on the search engine results page and relevancy of the page was investigated in [12]. The experiments showed that including cursor movements and scrolling was more effective than considering only the dwell time to estimate the page (document) relevancy. Similarly, in [13] the search performance was enhanced significantly using text-selection data. User post-click behaviour parameters such as mouse clicks, mouse movements, text selection and cursor trails are used to cluster the users based on their behaviour similarity as described in [14].

A document can be represented using the vocabulary used by the user during the retrieval of the document [17]. Recently, [18] integrated the content-based (TF-IDF) and the connectivity-based ranking algorithms using the click-through data to improve the search result for a web page. Another approach was proposed to develop a snippet-based algorithm to estimate the document relevance. The proposed algorithm was found to be more efficient than the commercial search engines [19].

Another post-click parameter which has been deemed to be useful is page review or re-finding. It is argued that about 30% of the user queries are used to retrieve a page which the user has previously visited [20]. Other research has proposed a page review based algorithm to predict the page relevancy and has shown the retrieval performance can be significantly improved [21].

Fuzzy logic systems (FLSs) have been applied to a range of application areas in Information Retrieval (IR) that include information filtering and personalised search. In this paper we are focusing on the approaches which used fuzzy logic to handle the uncertainty and subjectivity in the user feedback. The recursive method is a single individual fuzzy based recommending method [22] in which the recommendation is created recursively and based on the users profile without using any other collaborative preferences. The fuzzy sets were used to model the recommended object as well as justifying the recommendations. Cornelisa et al [23] proposed a fuzzy based conceptual framework for recommending one-and-only items. One-and-only items are the items which have only one occurrence in the data. The single occurrence of such items limits the classic collaborative filtering abilities to recommend the required item. The fuzzy logic was used for user preferences modelling to justify the similarity calculation.

Carbo and Molina [24] developed a collaborative filtering based algorithm in which the linguistic labels and the associated fuzzy sets were used to handle the uncertainty and inaccuracy in ranking the retrieved items. In [25] and [26] a hybrid fuzzy approach was proposed, to support the decision making process of individuals when seeking recommendations from other individuals about their personal selections. Here, fuzzy logic was used to model the similarity of an individual's feelings towards a specific item which in this case was a movie.

In [27] a fuzzy based agent to rank recommended candidates CVs within recruitment systems was proposed. Fuzzy logic was used to model the job preferences of the selection board members and also to resolve the uncertainty and conflict in the group decision making. A similar fuzzy based approach was proposed for activity-led learning [28]. A fuzzy based method that improves the collaborative filtering efficiency for multiple collaborating users was proposed in [29]. In this system fuzzy sets were used to model the user bias and uncertainty which result from multiple user interaction.

The existing approaches described above were mainly focused on methods identifying indicators of document relevancy and user preference. They did not consider combining these methods with the need to create profiles both of the user and the task, regarding the relevance of the returned information. Also, many of these approaches focus on well described content such as news stories, events, and movies and not on the unstructured content (documents), commonly found in enterprise systems, which have less descriptive details. In addition, such systems also have to contend with uncertainties (subjectivity and inconsistencies) in information relevance and the user perception of relevancy in respect of the retrieved content. Finally, these approaches do not integrate implicit and explicit feedback parameters with the query text analysis in order to gain more reliable relevance feedback.

III. PROPOSED METHOD

The relevance feedback including user query is used as the main data source for developing three profiles: task profile, user profile and document profile. The task profile is modelled as a sequence of weighted terms. The term weight reflects relevance level of the term to the task, the user and the document which the profile is related to. However, relevance feedback involves a high level of uncertainty due to inconsistency in user behaviour and subjectivity in their assessment of relevancy [30]. Therefore, handling such uncertainty is crucial to achieve better performance. We use a fuzzy approach to overcome

the uncertainty and bias in user judgment in order to provide a normalized ranking method for enterprise search. Our approach consists of six phases as shown in *Fig.1.*

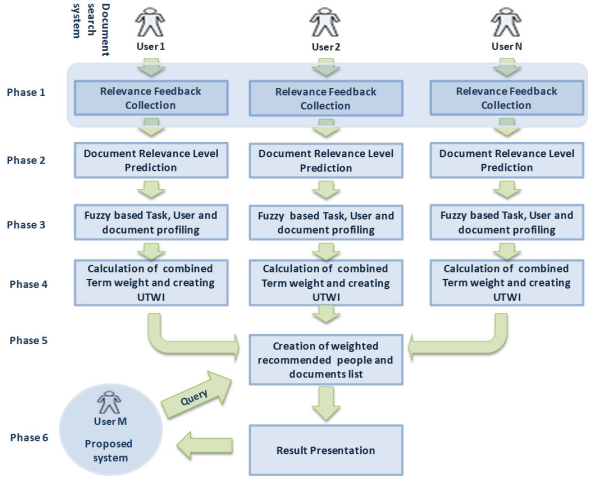


Figure. 1. The Proposed Approach

A. *Phase 1: Relevance Feedback Collection*

In this phase the relevance feedback is captured from the users during the search process. The captured relevance feedback includes implicit parameters, explicit parameters and user queries. The implicit parameters include: document Id, document hyperlink, visit time stamp, time on page, number of mouse clicks, mouse movement, mouse scrolling, scroll bar holding, key down times, key up times, book mark, save and print. Explicitly, the users are asked to rate the visited documents indicating their relevance to the query. The users and their tasks are identified through their unique user IDs.
The query information include: query text, query time stamp, number of retrieved documents based on the query.

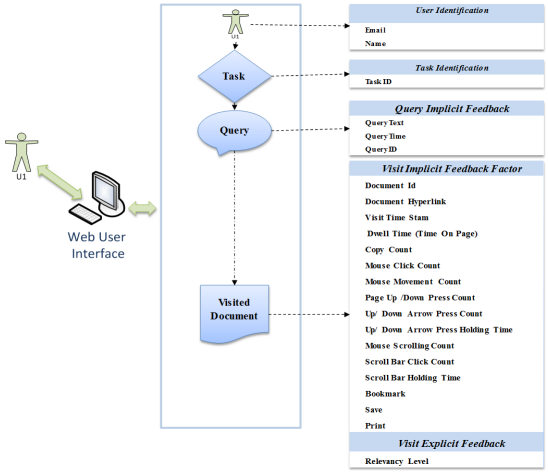


Figure. 2. Relevance Feedback Collection

B. *Phase 2: Document Relevance Prediction*

In this phase the relevance level of the visited documents is predicted from the implicit feedback parameters. The predicted value was calculated, as shown in TABLE I, using a linear predictive model based on linear regression analysis. The model was validated using *R*-squared (R^2) method which is a common accuracy validating method for regression models [31].The accuracy of the model is 76.5 %.

TABLE I. COEFFICIENTS FOR THE TARGET EXPLICIT RELEVANCE FEEDBACK

Model Term		*Coefficient* (β_i)		*Sig*	*Importance*
Intercept	(β_0)	1.395		.000	-
Time on Page	(X_1)	.0.069	(β_1)	.021	0.893
Mouse Scroll Count	(X_2)	0.013	(β_2)	.012	0.079
Mouse Movement Count	(X_3)	0.113	(β_3)	.031	0.028

C. *Phase 3: Fuzzy based Task, User and Document profiling*

In this phase, three types of profiles are created: the search task profile which is predefined and related to the role [32], the user profile and the document profile. The profiles are created by employing an adaptive fuzzy approach [33]. The approach contains 18 fuzzy rules to calculate the terms' fuzzy weights based on the term frequency measures NDF, NNTDF and NIDF. In this paper, we modified the approach to suit the query text analysis And to create the profiles through the following steps.

Step 1, we selected the set of queries $\boldsymbol{Q}$ which led to document visits. Subsets $\boldsymbol{\Omega_{Ocx}}$ of the query set $\boldsymbol{Q}$ are identified based on a collection $\boldsymbol{O}$ where $\boldsymbol{c}$ = {'*s*', '*u*','*d*'}, '$\boldsymbol{s}$' denotes task, '$\boldsymbol{u}$' denotes user and '$\boldsymbol{d}$' denotes document, and *x* is an identifier (referring to a particular task, user or document related queries) in $\boldsymbol{O_{cx}}$.

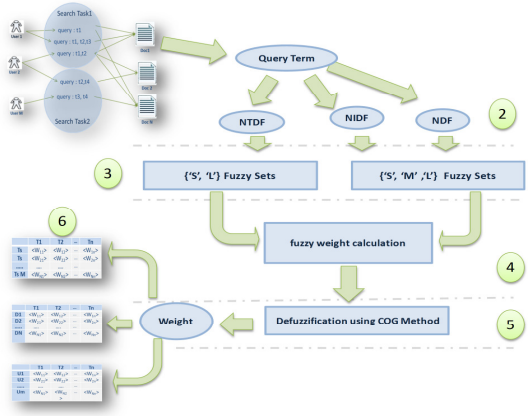


Figure. 3 Fuzzy based Task, User and document profiling

Step 2, after identifying the sets $\boldsymbol{\Omega_{Ocx}}$ in step 1, the queries in each set were pre-processed and transformed into a set of candidate terms through eliminating stop-words and stemming by Porter's algorithm [34]. The frequency measures: Distributed Term Frequency DTF, Document Frequency (DF), and Inverse Document Frequency (IDF) of each candidate term were calculated and normalized based on each set $\boldsymbol{\Omega_{Ocx}}$ and used as inputs to a fuzzy system for calculating a weight for each term.

These frequency measures are used to calculate the term frequency in a document collection. However, they

are also used in a collection of user queries where each user query could be considered as a document in order to calculate the frequency of the query terms as described in [35] and [16]. Based on that, in this step only, both terms: '*document*' and '*query*' refer to the user query. The DTF reflects the frequency and distributed status of a term in a set of user queries. This is calculated by dividing total occurrences of the term in the query set $\boldsymbol{\Omega_{Ocx}}$ by the number of the queries which contain the term in the set $\boldsymbol{\Omega_{Ocx}}$. The DF represents the frequency of queries having a specific term within the set $\boldsymbol{Q}$. The Normalized Document Term Frequency (NTDF), is defined in (1).

$$NDTFi = \frac{\frac{TF_i}{DF_i}}{Max_j\left[\frac{TF_j}{DF_j}\right]} \quad (1)$$

Where, $\boldsymbol{TF_i}$ is the frequency of term $\boldsymbol{t_i}$ in the query set $\boldsymbol{\Omega_{Ocx}}$, $\boldsymbol{DF_i}$ is the number of queries having term $\boldsymbol{t_i}$ in the query set $\boldsymbol{\Omega_{Ocx}}$. $\boldsymbol{i}$ and $\boldsymbol{j}$ = 1 to M where M is the number of the terms in the set $\boldsymbol{\Omega_{Ocx}}$.

The Normalized Document Frequency (NDF), is defined in (2).

$$NDFi = \frac{DF_i}{Max_j\, DF_j} \quad (2)$$

Where; $\boldsymbol{DF_i}$ is the number of queries having term $\boldsymbol{t_i}$ in the in the query set $\boldsymbol{\Omega_{Ocx}}$.

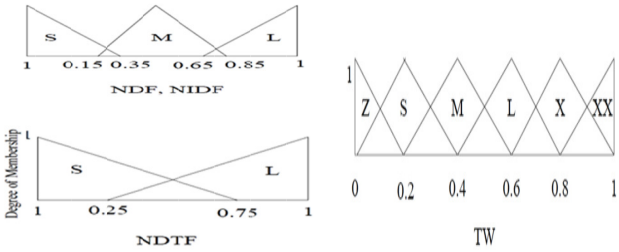
The IDF represents the frequency of the term in the query set $\boldsymbol{Q}$ rather than the set $\boldsymbol{\Omega_{Ocx}}$. We used IDF to identify the terms which appear in many queries which might relate to different tasks, users and documents. These terms are not very useful for representing the relevance level and consequently they will be given a less weight than the others. The Normalized Inverse Document Frequency (NIDF) is defined as follows:

$$NIDFi = \frac{IDF_i}{Max_j\, IDF_j}, \qquad IDFi = Log\ \frac{N}{n_i} \quad (3)$$

Where, N is the total number of queries in Q and $\boldsymbol{n_i}$ is the number of queries in Q in which the term $\boldsymbol{t_i}$ appears.

Step 3, in this step, the crisp values of the three input variables (NDTF, NDF, and NIDF) are fuzzified and mapped to sets of predefined fuzzy sets. As shown in *Fig. 4. a*, NDTF and NIDF have three linguistic labels { S(Small), M(Middle), L(Large) }, and NTDF as has two linguistic labels { S(Small), L(Large) }. As shown in *Fig. 4. b,* the output variable *TW* has six fuzzy sets associated with six linguistic labels {Z(Zero), S(Small), M(Middle), L(Large), X(Xlarge), XX(XXlarge) }.

Step 4, in this step, we used the 18 '***If →Then***' fuzzy rules which are described in [33] to infer a fuzzy term weight (*TW*) for the term $\boldsymbol{t_i}$. These rules are constructed based on the assumption that the important or representative terms may occur across many queries in the representative query set $\boldsymbol{\Omega_{Ocx}}$ but not in the whole selected query set Q. In other words these terms have high NDF and NIDF values and low NTDF values. For example, as shown in *Fig. 5*, when NDF of a term is high and its NIDF is also high, the term is considered as a representative keyword so the output weight is between X and XX.



a).Input Variables b). Output Variable *TW*

Figure. 4. Fuzzy Sets for Input Variables

NIDF / NDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NDTF = S

NIDF / NDF	S	M	L
S	Z	S	M
M	Z	L	X
L	S	X	XX

NDTF = L

Figure. 5. *WT* Calculation Fuzzy Rules

Step 5, the output of step 4 *TW* is defuzzified using the center of gravity (COG) method in order get a crisp weight TW_{t_i} for each term to be added to the profile associated with the collection $\boldsymbol{O_{cx}}$.

Step 6, in this step the term $\boldsymbol{t_i}$ with its weight TW_{t_i} are added to the profile being created. However, as the system is used, more relevance feedback, including user queries will be captured and the system will calculate new weights if the term frequencies change. The profile will then be updated by changing the term weight(s) to the new value(s).

More formally, let's assume P_{Ocx} is the profile associated with the collection $\boldsymbol{\Omega_{Ocx}}$ and M is the number of terms in $\boldsymbol{\Omega_{Ocx}}$ then the profile $\boldsymbol{P_{Ocx}}$ is defined as a set of weighted terms as follows:

$$P_{O_{cx}} = \bigcup_{i=1}^{M} t_i TW_{t_i} \quad (4)$$

As described in the previous steps, this phase results in creating/updating profiles for the tasks, the users and the documents in the data set. These profiles are combined into one index in the next phase.

D. Phase 4: Fuzzy combined Weight calculation

In this phase the task, user and document profiles are combined in one index which is called the Unified Term Weight Index (UTWI). In this index each term has a unified weight per task, per user and per document. If term $\boldsymbol{t_i}$ was used by user $\boldsymbol{U_k}$ to retrieve the document $\boldsymbol{D_g}$ in order complete the search task $\boldsymbol{Sy}$ then the unified term weight for $\boldsymbol{t_i}$ is $\boldsymbol{W_{iykg}}$. This means that the new weight considers the relevance between the term and the whole combination of the three factors; the user, the document and the task. This phase includes the following steps:

Step1, *Fuzzy rules Extraction*, let's say that *V* is the set of document visits in the data set which contains *H* visits, then V_h is the document visit instance where $h=$ 1 to *H*. Each V_h is associated with the user query Q_e which led to this visit where e=1 to E and E is the number of queries in the dataset, the search task it occurred in S_y, the user who made this visit U_k, the visited document D_g and the predicted relevance feedback R_h (see output of phase 2). Q_e consists of Z terms where t_{ez} is the query term in Q_e and z =1 to Z. Then each t_{ez} is associated with its weight W in each of the profiles of S_y, U_k, and D_g that were computed in phase 3. These three weights are associated with the predicted relevance R_h. As a result, each term t_{ez} is represented as a set of four values $\{W_{s_y t_{mz}}, W_{u_{yk} t_{mz}}, W_{d_{yg} t_{mz}}, R_h\}$. If we consider the three first weights as inputs and the R_h as a result, then we have a sequence of three input values and one result value $\{W_{s_y t_{mz}}, W_{u_k t_{mz}}, W_{d_g t_{mz}} \rightarrow R_h\}$ for each instance in dataset.

The inputs and output values are mapped to predefined fuzzy sets with the linguistic labels 'Low' (L), 'Medium' (M) and 'High' (H) based on Mendel Wang method described in [7] as shown in *Fig 6* In our system, the shapes of the membership functions for each fuzzy set are based on triangle MFs as shown in *Fig. 6*. The outcome from this step is a set of antecedents and consequents also called '***if →then'*** fuzzy rules where each of the inputs and the outputs are represented by the associated linguistic label as shown in TABLE II. If B is the linguistic label {'L','M','H'} of the value of each of the inputs and the output then the fuzzy rule FR_h is:

$$B\left(W_{s_y t_{mz}}\right), B\left(W_{u_k t_{mz}}\right), B\left(W_{d_g t_{mz}}\right) \rightarrow B(R_h) \qquad (5)$$

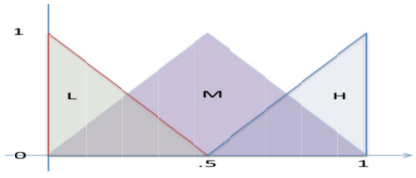


Figure. 6. Fuzzy Sets for Input and Output Variables

Step 2, *Compression of Fuzzy Rules*, in this step, we performed a rule compression on the fuzzy rules resulting from the previous step. This was done in order to extract those rules with the maximum firing strength. This process involves a modified calculation of two rule quality measures from which we then derive the scaled fuzzy weight of each unique summarisation rule. The quality measures are based on generality which measures how many data instances support each rule and reliability that measures the confidence level in the data supporting each rule [7].

In our approach the rule generality is measured using scaled fuzzy support and the reliability of the rule is based on its scaled confidence level. The fuzzy support of a rule is calculated as the product of the rule's support and firing strength. The support of a rule refers to coverage of data patterns that map to it, while its' firing strength measures the degree to which the rule matches those input patterns.

TABLE II. SAMPLE OF THE EXTRACTED FUZZY RULES

User	Task	Document	Term	W_u	W_T	W_D	→	R
U1	T2	11884897.html	PUBLICATION	H	M	M	→	L
U2	T20	09530858.html	PUBLICATION	M	M	M	→	H
U3	T3	15292585.html	DIFFER	M	M	M	→	M
U4	T3	15292585.html	SHEET	M	M	H	→	M
U5	T3	15292585.html	TRIAL	M	M	M	→	M
U3	T6	01314419.html	DIETARY	M	M	H	→	L
U5	T2	01314419.html	TOTAL	L	M	M	→	H

The rule's fuzzy support can be used to identify the unique rules with the most frequent occurrences of data patterns associated with them, where the data patterns also most closely map to those rules. The fuzzy support of each rule is scaled based on the total data patterns for each output set so that the frequencies are scaled in proportion to the number data patterns found in each consequent set. The calculation of the scaled fuzzy support for a given uniquely occurring rule is shown in (6) and is based on the calculation described in [36]. In our approach it is used to identify and eliminates duplicate instances by compressing the rule base into a set of M unique and contradictory rules modelling the data.

$$scFuzzSup\left(\underline{FR_l}\right) = \frac{Co_{\underline{FR_l}}}{Co_{\underline{FR_l}} + Co_{\widehat{FR_l}}} \qquad (6)$$

Where l =1 to M , l is the index of the rule $\underline{FR_l}$ is a unique antecedent combination associated with the consequent linguistic label B and $Co_{\underline{FR_l}}$ is the number of instances which support the rule $\underline{FR_l}$ in the data set. $\widehat{FR_l}$ is the set of contradictory antecedents' combinations (the other antecedents combination which are different to $\underline{FR_l}$ but have the same consequent as of $\underline{FR_l}$) $Co_{\widehat{FR_l}}$ is the number of the instances which support these other combinations $\widehat{FR_l}$.

The confidence of a rule is a measure of a rule's validity describing how tightly data patterns are associated to a specific output set. The confidence value is between 0 and 1. A confidence of 1 means that the pattern described in the rule is completely unique to a single output set. A confidence of less than 1 means that the pattern described in the rule occurs with more than one output set, and would then be associated with the output set with the highest confidence. The rule scaled confidence calculation is shown in (7) and is based on the calculation described in [36].

$$scConf\left(\underline{FR_l}\right) = \frac{scFuzzSup\left(\underline{FR_l}\right)}{Co_{\widehat{RF}_l}} \qquad (7)$$

Step3, *Calculation of Scaled Rule Weights,* In this step, the product of the scaled fuzzy support and confidence of a rule is used to calculate the rule's scaled fuzzy weight as shown in (8).

$$scWi = scFuzzSup \times scConf \qquad (8)$$

Each of the generated M rules is assigned the scaled fuzzy weight measure ***scWi*** and takes the following form:

$$B\left(W_{s_y t_{mz}}\right), B\left(W_{u_k t_{mz}}\right), B\left(W_{d_g t_{mz}}\right) \rightarrow B(R_h)[scWi] \quad (9)$$

The scaled fuzzy weight measures the quality of each rule in modelling the data. It can be used to rank the top rules associated with each output set and to choose a single winner rule among compatible rules, based on methods for rule weight specification described in [36]. We used these weights to extract the most representative rule patterns where the pattern with the highest value of ***scWi*** was selected over the other contradictive patterns. The selected patterns were used in a fuzzy system, as described in the following step, for modelling the relevancies based on the most important profile weighted terms.

Step 4, *Calculation of the unified term weight,* in this step the resulting rules from the previous step are used to build a fuzzy system to calculate the unified term weight $\boldsymbol{W_{iykg}}$ for each query term $\boldsymbol{t_i}$ in each associated document visit $\boldsymbol{V_h}$. The fuzzy system calculates the unified term weight based on the term weights in the profiles of the associated user $\boldsymbol{U_k}$, document $\boldsymbol{D_g}$ and search task $\boldsymbol{S_y}$ which were created in phase 2 and the fuzzy rules extracted in step 3 of this phase.

The fuzzy system consists of the three input variables { $\boldsymbol{W_{s_y t_{mz}}}$, $\boldsymbol{W_{u_k t_{mz}}}$, $\boldsymbol{W_{d_g t_{mz}}}$} ,one output variable which is the unified term weight $\boldsymbol{W_{iykg}}$, and the fuzzy rules which are extracted in step 3. The fuzzy system is then fed with the values of the inputs: $\boldsymbol{W_{s_y t_{mz}}}$, $\boldsymbol{W_{u_k t_{mz}}}$and $\boldsymbol{W_{d_g t_{mz}}}$ which were associated with each query term for each document visit in step 1. The calculated value of $\boldsymbol{W_{iykg}}$ was then was used to create the UTWI which consists of { $\boldsymbol{V_h}$,$\boldsymbol{t_i}$, $\boldsymbol{S_y}$, $\boldsymbol{U_k}$, $\boldsymbol{D_g}$, $\boldsymbol{W_{iykg}}$}. UTWI was used in the next phase to create the recommendations.

E. Phase 5: Recommendation of documents and people.

In this phase the proposed system's user query is pre-processed in the same way as in phase 1.

The UTWI index is searched for the extracted query terms to find matching documents and people who visited those documents frequently. This starts with finding the matching tasks in order to recommend documents and people based on the relevant task. A matching task should have at least one occurrence of at least one of the query terms in its associated instances in UTWI. Then for each of these tasks the averages of the matching terms' weights are calculated. Then these weights are summed to give the aggregate task weight. Based on the aggregate task weight the relevant documents and users of each matching task can be extracted. A relevant document/user should have at least one matching term occurrence in UTWI with the tasks terms. The average weight of each matching term is calculated for the relevant document/user and these are summed to calculate the aggregate weight of the relevant document/user. The document/users are sorted in descending order based on their aggregate weights.

F. Phase 6: Recommendation Presentation

In this phase, the recommended document and search users are presented through a web-based graphical user interface. Through this interface the recommended documents can be viewed as a weighted list where the relevance weight of each document to the query is shown. Similarly, it also shows a user analysis chart containing the relevant search user with their relevance weight to the query. The search queries made by each person are shown, together with the relevance weight of the information returned by the query. The interface provides a query-task tree in which each query is displayed. For each query the people making that query on a task are grouped according to the relevance given by the system to the data returned by that query.

IV. EXPERIMENTS AND RESULTS

A. Experimental Set up

In order to run our experiments, we have selected a TREC Enterprise 2007 Track, a slandered test collection for enterprise search [37]. The data set was labelled as the test collection and provided a group of 50 queries which were previously created by real users and associated with the relevant documents for each query according to users' judgment [37]. The data set was then extended by creating search tasks. We invited 35 users to participate in the experiment. The users were asked to freely formulate their queries in order to search for information which can help them to find solutions to those tasks.

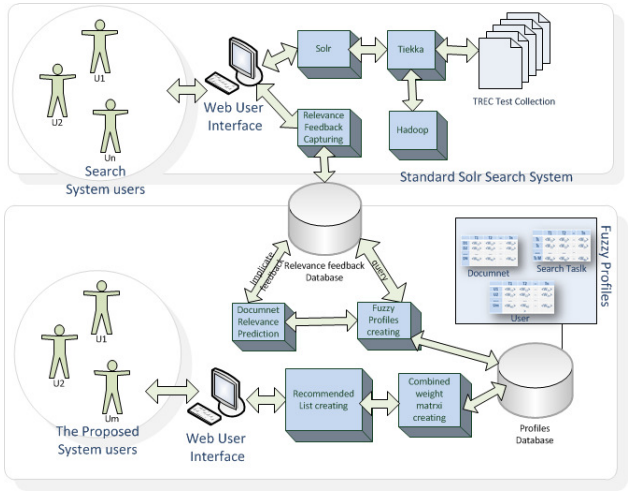


Figure. 7. Experimental Set Up

After preparing the search tasks, *TREC Enterprise 2007 Track* was indexed using a configured text based search system. The configured system is based on the open source technology and consists of the following components: Apache Solr, Apache Tika and Hadoop. Hadoop is an open source framework for distributed computing. Tika is an open source toolkit that can parse and acquire different types of documents. Solr is an open source enterprise search server which is based on the underlying search library Lucene that is widely used in information retrieval applications [38]. A web user

interface was developed and integrated with the system to enable the users to search the corpus documents and also to capture the relevance feedback.

As discussed in *Section. II – Phase 1*. The system harvested 812 user queries and 1230 document visits which gave a reasonable size of relevance feedback for creating the search task, user, and document profiles. The captured implicit and explicit parameters values were used to develop the predictive linear model. Then, the captured users' queries were pre-processed using Oracle Text Search library. The extracted query terms were fed into the fuzzy system as discussed in *Section. II – Phase 3*. The resulting profiles were projected and associated with predicted document relevance level from the linear model. The resulting dataset instances were then mapped to linguistic labels in order to extract the fuzzy rules. The fuzzy rules were summarised as discussed in *Section. II – Phase 4*. The summarised fuzzy rules which are shown in TABLE III were used to build a fuzzy inference system in order to calculate the combined term weight as discussed in the same section.

TABLE III. SAMPLE OF THE SUMMARISED WEIGHTED FUZZY RULES

W_T	W_U	W_D		W_R	$scWi$
H	H	H	→	H	0.489292903
H	H	M	→	H	0.364747958
H	M	H	→	H	0.302517053
H	L	H	→	M	0.116836792
H	L	M	→	M	0.082494544
H	M	L	→	M	0.010468469
	…	…	…	..	…..

A web based user interface was developed to handle the users' queries as discussed in *Section. II – Phase 5*.

B. *Results and Evaluation*

The proposed approach was evaluated at two levels; the method validation and the overall system performance evaluation.

Fuzzy Rule Extraction and Summarisation Validation; Fuzzy Rule Extraction and Summarisation was validated using the Hold Out method as described in [39]. The rule extraction and summarising components were trained on 80% of the data set and then it was tested on the unseen 20% of the data set. The resulted rules were used to classify the relevancy of each instance in the unseen data as described in step 4 of phase 4. The resulted relevancy classifications were compared with the associated linguistic labels of the actual explicit relevance feedback values as shown in TABLE IV. These linguistic labels used the same fuzzy sets as in step 1 of phase 4. The resulting accuracy of our proposed system was 86%.

Overall Performance Evaluation, Precision ***(P)*** and Recall ***(R)*** are standard evaluation metrics used in information retrieval research [40]. We carried out a comparative performance evaluation in which we compared the values of ***P*** and ***R*** for the text based search system and the proposed system. As shown in the *Fig .8*, the average ***P*** value for text based search system is (0.011) which is relatively low. This low value ***P*** indicates that the system retrieves a large number of irrelevant documents. The proposed approach enhanced the value of ***P*** significantly where the average of ***P*** value rose from (0.011) to (0.064). The proposed system also enhanced the value of ***R***. As shown in *Fig .8,* the average value of ***R*** increased significantly from (0.436) to (0.828) which means the ability of the system to retrieve the relevant document is enhanced as well.

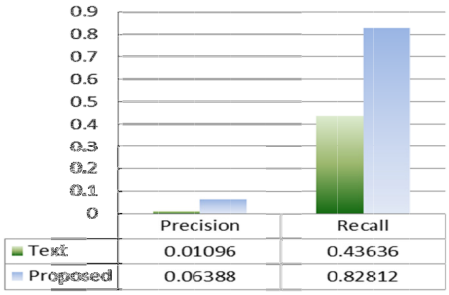


Figure.8. Average Precession & Recall Comparison

TABLE IV. SAMPLE OF SUMMARISED FUZZY RULES ACCURACY

	Task	User	Document	E	A	C
Cotton	T13	**@coventry.ac.uk	04574741.html	M	M	1
Tech	T1	**@coventry.ac.uk	04587909.html	M	L	0
Air	T12	**@uni.coventry.ac	04736857.html	M	M	1
Guitar	T13	**@yahoo.com	12228999.html	M	M	1
Australia	T12	**@coventry.ac.uk	02493670.html	M	M	1
Australia	T13	**@gmail.com	03007618.html	M	M	1
School	T1	**@coventry.ac.uk	16400222.html	M	L	0
Cooper	T13	**@coventry.ac.uk	16400222.html	M	M	1
Cooper	T13	**@yahoo.com	16400222.html	M	M	1
Bio	T12	**@yahoo.com	16456196.html	M	M	1
Lab	T13	**@gmail.com	08225989.html	M	M	1

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an approach for the development of a fuzzy IR system. The approach provides a new mechanism for constructing and integrating three relevancy profiles: task profile, user profile and document profile into a unified index though the use of relevance feedback and fuzzy rule based summarisation. The fuzzy approach was used to create the profiles and integrate them. The motivation for using the fuzzy approach was to handle the uncertainty due to inconsistency and subjectivity in the users' assessment of relevance feedback.

We ran experiments in which the relevance feedback was captured from 35 users on 20 predefined simulated enterprise search tasks. During this process the system captured implicit and explicit feedback parameters, and the user queries. The captured dataset was used to develop and train the fuzzy system. The system showed an 86% performance accuracy in correctly classifying document relevance. The overall performance of the proposed system was evaluated based on standard precision and recall which showed significant improvements in retrieving relevant documents. Our future work will include further evaluation of the proposed approach in a real world organisation using an extended user base over a longer timescale.

VI. RERFRENCES

[1]. S. Feldman, C. Sherman, " The high cost of not finding information," IDC technical report, April 2008.

[2]. Hawking, R. Baeza-Yates, B. Ribeiro-Neto ," Modern Information Retrieval," 2nd ed. Pearson Educational, , pp. 641-68, 2010.

[3]. P. C. Vaz , D. de Matos , B. Martins, "Stylometric relevance-feedback towards a hybrid book recommendation algorithm, Proceedings of the fifth ACM workshop on Research advances in large digital book repositories and complementary media," Maui, Hawaii, USA, October, 2012

[4]. G. Jawaheer , M. Szomszor , P. Kostkova," Comparison of implicit and explicit feedback from an online music recommendation service," Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, p.47-51, Barcelona, Spain, Sept 2010.

[5]. S. Schiaffino , A. Amandi," Intelligent user profiling, "Artificial intelligence: an international perspective," Springer-Verlag, pp. 193-216, Berlin, 2009.

[6]. G. Adomavicius, A. Tuzhilin, " Toward the next generation of recommender systems: a survey of state-of-the-art and possible extension, " IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, 2005.

[7]. Wu, J. M. Mendel, and J. Joo, "Linguistic Summarization Using If-Then Rules," Proceedings of the IEEE International Conference on Fuzzy Systems, Barcelona, Spain, July, 1-8, 2010.

[8]. S. Jung, J. Herlocker, J. Webster, " Click data as implicit relevance feedback in web search," Information Processing & Management, Vol. 43, pp. 791–807, 2007.

[9]. M. Morita, Y. Shinoda, " Information ltering based on user behavior analysis and best match text retrieval, "Proceedings of the 17th annual international ACM SIGIR conference, Springer-Verlag New York, pp. 272-281,1994.

[10]. J. Kim , D. Oard,. K, Romanik, " Using implicit feedback for user modeling in internet and intranet searching," University of Maryland CLIS Technical Report 00-01. 2000., 2001.

[11]. A. Guo, E. Agichtein," Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior," Proceedings of the 21st international conference on World Wide Web, pp. 569–578, 2012.

[12]. R. White, & G. Buscher, " Text selections as implicit relevance feedback," InProceedings of the 35th international ACM SIGIR conference on research and development in, information retrieval pp. 1151–1152, 2012.

[13]. G. Buscher, W. White, S. Dumais, J. Huang, "Large-scale analysis of individual and task differences in search result page examination strategies," Proceedings of the fifth ACM international conference on web search and data mining, pp. 373–383, 2012.

[14]. V. Balakrishnan, X. Zhang, "Implicit User Behaviours to Improve Post-Retrieval Document Relevancy," Computer in Human Behavior, Vol. 33, pp. 104–112, 2014.

[15]. M. Claypool, D. Brown, P. Le, M, Waseda, "Inferring User Interest, " IEEE:Internet Computing, pp. 32-39, 2001

[16]. B. Poblete , R. Baeza-Yates, "Query-sets: using implicit feedback and query patterns to organize web documents," Proceedings of the 17th international conference on World Wide Web, Beijing, China, April 2008.

[17]. B. Bidokia, P. Ghodsniab, N. Yazdanic, F. Oroumchiand, "A3CRank: An adaptive ranking method based on connectivity, content and click-through data," Information Processing & Management, Vol. 46, no. 2, pp. 159–169, 2010.

[18]. Y. Liu, J. Miao, M. Zhang, S. Ma, L. Ru, "How do users describe their information need: Query recommendation based on snippet click model," Expert Systems with Applications, vol. 38, no. 11, pp. 13847–13856, 2011.

[19]. S. Tyler, J. Teevan, " Large scale query log analysis of re-finding," In Proceedings of the third ACM international conference on web search and data mining, pp. 191–200, 2010.

[20]. S. Tyler, J. Wang, Y. Zhang, "Utilizing re-finding for personalized information retrieval," Proceedings of the 19th ACM international conference on information and knowledge management , pp. 1469–1472, 2010.

[21]. V. Balakrishnan, X. Zhang, "Implicit User Behaviours to Improve Post-Retrieval Document Relevancy," Computer in Human Behavior, Vol. 33, pp 104–112, 2014.

[22]. R. Yager, "Fuzzy logic methods in recommender systems", "Elsevier Science B.V. Fuzzy Sets and Systems", vol.136, pp.133 – 149, 2003.

[23]. Cornelisa, J. Lub, , X. Guob, , G. Zhangb, "One-and-only item recommendation with fuzzy, " Information Sciences, vol. 177, No. 22, pp. 4906–4921, November 2007.

[24]. J. Carbo, J.M. Molina, " Agent-based collaborative filtering based on fuzzy recommendations, " Int. J. Web Engineering Technology, vol. 1, No. 4, pp. 414–426, 2004.

[25]. P. Perny, J.D. Zucker, "Preference-based search and machine learning for collaborative filtering: the 'Film-Conseil' movie recommender system, " Revue, vol. 1, No. 1, pp. 1–40, 200.

[26]. L, Campos , J. Fernández-Luna , J. Huete, "A collaborative recommender system based on probabilistic inference from fuzzy observations," Fuzzy Sets and Systems, v.159, No .12, pp .1554-1576, June, 2008.

[27]. F. Doctor, D. Roberts, V. Callaghan, "A Fuzzy Based Agent for Group Decision Support of Applicants Ranking within Recruitment Systems," Proceedings of the Intelligent Agents,2009. IA '09. IEEE Symposium on, pp. 8-15, USA, Nashville , 2009.

[28]. R. Iqbal, F. Doctor, N. Shah, X. Fei, "An intelligent framework for activity led learning in network planning and management," Journal of Computer Networks and Distributed Systems, 12(4), pp .401-419, Inderscience, 2014.

[29]. A. Eckhardt, "Similarity Of Users' (Content-Based) Preference Models For Collaborative Filtering In Few Ratings Scenario, " Expert Systems with Applications, vol .39, No. 14, pp. 11511–11516, October 2012.

[30]. A. Grzywaczewski, R. Iqbal, "Task-Specific Information Retrieval Systems for Software Engineers," Journal of Computer and System Sciences, vol. 78, no. 4, pp. 1204-1218, 2012.

[31]. D. Campbell, S. Campbell, "Introduction to Regression and Data Analysis," Statlab Workshop, October 28, 2008.

[32]. F. Hoque, "e-Enterprise: business models, architecture, and components, " ed. 1, Vol. 2, pp. Cambridge University Press, 2000

[33]. Q. Li, B. M. Kim, "Constructing User Profiles for Collaborative Recommender System," Advanced Web Technologies and Applications Lecture Notes in Computer Science, vol. 3007, pp. 100-110, 2004.

[34]. M.F. Porter, "An algorithm for suffix stripping, " Program, vol. 14, no. 3, pp.130 – 137, 1980.

[35]. M. Castellanos, " Hotminer : Discovering Hot Topics Formality Text," Survey of Text Mining. Springer-Verlag New York, vol. II, pp 123-157,2003.

[36]. H. Ishibuchi and T. Yamamoto, "Rule Weight Specification in Fuzzy Rule-Based Classification Systems," IEEE Transactions on Fuzzy Systems, vol. 13, no. 4, pp. 428–435. August 2005.

[37]. P. Bailey, N. Craswell, A. P. Vries, I. Soboro, " Overview of the TREC-2007 enterprise track (2007)," Proceedings of the Fourteenth Text REtrieval Conference (TREC), pp. 24-33, 2008.

[38]. O. Alhabashneh, R. Iqbal, N. Shah, S. Amin, A. James, "Towards the development of an integrated framework for enhancing enterprise search using latent semantic indexing," Proceedings of 19th International Conference on Conceptual Structures, pp. 346-352, Derby, UK, July, 2011.

[39]. S. Arlot , A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, pp. 40–79, 2010.

[40]. D. Kelly, "Methods for Evaluating Interactive Information Retrieval Systems with Users, " Foundations and Trends in Information Retrieval, vol. 3, no. 1–2 , pp. 200–224, 2008.