Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110526

User Profile based Information Retrieval Incorporated with Reinforcement Learning

S. Subitha[1,*] *and S. Sujatha*[2]

[1] M.A.M College of Engineering, Siraganur, Tiruchirappalli, Tamilnadu, India.
[2] Bharathidasan Institute of Technology, Anna University, Tirchirappalli, Tamilnadu, India.

Received: 6 Jul. 2017, Revised: 20 Aug. 2017, Accepted: 25 Aug. 2017
Published online: 1 Sep. 2017

Abstract: Information retrieval is a complex process that involves understanding the user's requirements to provide appropriate and relevant results. This involves combined working of several techniques such as contextual analysis, correlation analysis, sentiment analysis and a good understanding of the user's profile. This paper presents an effective relevance feedback based information retrieval model that aids in effective retrieval and organization of results such that information relevant to the users are given high priority. The user's profile is constructed and reinforced with their queries and selection responses. This is iteratively performed such that the user's profile gets strengthened with better and more appropriate rules. Result organization is performed based on the significance levels, sentiment and user's preferences. Experiments on STS Gold Sentiment Corpus indicate effective predictions when compared with recent models.

Keywords: Information Retrieval, Context, Sentiment Analysis, Data Significance, User Profiling

1 Introduction

Exponential growth of information available online has led to the increase in online data. This, although is an advantage, it has its own side effects. The huge amount of information pertaining to all the domains has led to an information cloud, from which filtering out the necessary information becomes difficult and tedious. Moreover, web being an amalgamation of information, has data pertaining to several domains. A single retrieval contains results from several domains leading to huge data and data conflicts. Due to the absence of a regulatory source in the internet, it can also contain wrong information and sometimes conflicting and outdated information. These issues serve as the major drawbacks of during the process of information retrieval. Fuzzy queries results in large amounts of data being retrieved from the repository, while most of the results are not related to the user's requirements [1]. The retrieved results usually contains mixed domains. User's requirements can be identified only by understanding the context or domain under which the query was posted upon. Identifying such components requires knowledge of the user's profile. User profile building refers to identifying the user's interests and domains of relevance. This can enable effective domain based filtering, hence higher correlation towards the user's requirements [2]. Learning the user's interests is not a standalone and a single level learning process. It requires continuous user activity monitoring and recording. It can also be made robust based on user's feedbacks. Apart from identifying domains, identifying a user's interest also requires their sentiments towards topics in the domains [3]. This requires identifying the sentiments of queries and interested results. Maintaining such user based information leads to fine-tuned and highly relevant results.

Several contributions related to content retrieval and sentiment analysis can be found in literature. Some of the recent researches in this domain are discussed here. Deep learning based models dealing with cross-media retrieval is proposed by Jiang et al. in [4]. This models deals with content retrieval from the internet. However, the proposed content is not limited to text; it is further extended to image based retrieval. Other basic content retrieval models include; a cross-modal multimedia retrieval model by Rasiwasia et al. in [5] and a multi-view clustering model by Chaudhuri et al. in [6]. Social networks have become the current platforms for information collection,

[*] Corresponding author e-mail: subitha.haran@gmail.com sujathaaut@gmail.com

© 2017 NSP
Natural Sciences Publishing Cor.

due to their high adoption levels. A survey on how social networks impact the information retrieval process is presented by Bouadjenek et al. in [7]. Several non-personalized information retrieval models have been proposed specifically for operating on social networking data. A logical inference based query expansion model was proposed by Lioma et al. in [8]. A similar model was proposed by Jin et al. in [9], incorporating social tags as the basic profile components. Tag based matching or similarity infers high correlation. A unified framework to address complex queries was presented by Mantrach et al. in [10]. A term enrichment based model was proposed by Lin et al. in [11].

Cross Lingual Information Retrieval (CLIR) and Multi-Lingual Information Retrieval (MLIR) models are also on the raise due to the increase in online documents in a specific language pertaining to a specific region. CLIR enables users to retrieve related documents that are in other languages and are not in the same language as the input query [12]. A review on MLIR on Indian languages was presented by Madankar et al. in [13]. A CLIR and GPS based data retrieval mechanism for effective data retrieval was proposed by Sharma et al. in [14]. A specific Marathi based CLIR system was proposed by Savitha et al. in [15]. Similarity based document retrieval models are currently being developed in-order to incorporate relevancy levels to the results obtained from user's query. A CASISS model for similarity identification during information retrieval in semi-structured documents was presented by Guezouli et al. in [16]. A knowledge based similarity identification model for information retrieval was proposed by Burke et al. in [17]. This model uses semi-structured documents to build their knowledge base for an FAQ finder tool.

Sentiment analysis also plays a vital role in identifying the relevancy of the document towards a user's query. However, it is not usually considered as a parameter during the information retrieval process. An emoticon based sentiment identification model was proposed by Nirmal et al. in [18]. This method is based on social networking data and operates on the text and the emoticons to determine the sentiment levels of the text. A customer preference mining model based on sentiment analysis was presented by Zhou et al. in [19]. This model is based on product reviews and uses the input query sentiment to retrieve appropriate content for the users. Similar such models include datamining review analysis by Hariri et al. in [20], investment trend based model by Bollen et al. in [21] and a general inquiry system by Stone et al. in [22].

This paper presents an architecture to effectively identify and retrieve appropriate results according to the user's requirements. The profile builder iteratively build the user's profile depending on their query and selection behaviors. The input query is analyzed and significant terms from the query are recorded in the user's profile. Results from query application are processed, filtered and sorted according to the user's profile. Finally user's feedback is incorporated into the profile for reinforced learning.

2 Relevance Feedback based Information Retrieval

Information retrieval operates on a broad domain, as the information being retrieved can be any data from the vast available information existing in the web. Information relevancy plays a major role in the retrieval process. Identifying the relevant information is not possible without monitoring and recording the user's access patterns [23]. User profile building plays a major role in defining the relevancy of the results. This paper proposes a relevance feedback based information ranking model to place appropriate and most relevant information at the top, hence improving the user's convenience. The relevance feedback based information ranking model is performed in four phases; user query processing, analysis of results retrieved from the web, profile based result ranking and profile update with reinforcement. The two major components of the proposed model are the query and result processing module and the user profile builder. Figure 1 shows the architecture of the proposed information retrieval model.

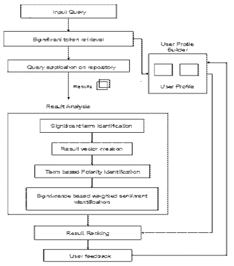
Input Query
Significant token retrieval
Query application on repository
Results
User Profile Builder
User Profile
Result Analysis
Significant term identification
Result vector creation
Term based Polarity Identification
Significance based weighted sentiment identification
Result Ranking
User feedback

Fig. 1: Feedback based Information Retrieval Architecture

2.1 User Profile Builder

User profile builder aids in building a customized repository for the user, which aids in providing the most relevant results. User profile is iteratively built for every query the user posts and every result the user examines from the list of results. The structure of a user's profile is shown below

The user's profile records the terms used for searching and the significant terms in the results preferred by the user. The user's preference towards a term might not be always biased towards a specific polarity. Hence the contextual polarity of the search result is also maintained in the profile. Finally, the frequency of preference of the

Table 1: User Profile

Term	Polarity	Frequency
Users query term/ Significant term in the result	Polarity context under which the term was specified	Number of occurrence

user for a particular term with a specific polarity is recorded. Hence a single term can have a maximum of three occurrences in the user profile table; term with positive polarity, term with negative polarity and term specified in the neutral context with their corresponding frequency levels. Hence when a query term is specified or when a result is selected, the corresponding frequency levels are updated.

2.2 Query and Result Processing Module

The query and result processing module operates on both, the query and the results retrieved for the query. The input query is processed to build the user profile. The retrieved results are also processed and analyzed to identify their relevancy levels in accordance with the user's query and the user's profile. These relevancy levels are used to provide ordered and customized results to the user according to their preference levels.

2.2.1 User Query Processing and Profile Update

User provides the input query, which serves as the base for the profile building process. The user query can comprise of specific terms or lines of text. The proposed model has been designed to handle both types of queries. Stop-word elimination is applied on the input query and the input tokens are added to the user profile builder. Sentiment level related to the input query is identified and the term along with the sentiment level are used to update the user's profile. A new entry is added to the user's profile if the entry is absent, otherwise, frequency of the available entry is incremented.

2.2.2 Result Analysis and Sentiment Identification

The input query is applied on the web and the retrieved results are used for analysis. The analysis phase includes identifying the significant terms, vector creation, term based polarity identification and significance based weighing of vectors for result ranking.

Significant Term based Vector Creation

The retrieved results are usually in the form of text, hence identifying the appropriate and usable content is required for analysis and ranking. This is done by identifying the significant terms appropriate to the search domain. This is done by finding the Term Frequency-Inverted Document Frequency (TF-IDF) of the terms. Higher value for TF-IDF indicates higher significance of the word in the document. TF-IDF is calculated using eq 1.

$$tfidf(t,d,D) = tf(t,d) * idf(t,D) \tag{1}$$

Term Frequency (TF) and the Inverted Document Frequency (IDF) are calculated using (2) and (3) [24]

$$tf(t,d) = \frac{tf(t,d)}{*} tf(t,d) \tag{2}$$

where, f(t,d) refers to the number of times the word t is contained in the document d and count(w,d) refers to the total number of words contained in the document d.

$$idf(t,D) = log\frac{N}{|d \in D : t \in d|} \tag{3}$$

where, N is the total number of documents in the corpus, and $|d \in D : t \in d|$ is the number of documents that contains word t.

The results are tokenized and TF-IDF for each of the tokens is identified. TF-IDF signifies the importance of the term in the current context. However, the retrieved results can contain several determiners and connectors. Such terms do not have any significance in the text. Such terms also have a TF-IDF value associated to it. Even though the value is low, it has to be neutralized so that accurate significance levels can be obtained. Hence a user defined threshold is applied on the list and all the terms with values below the defined threshold are set to zero. Result vectors are created with the significant terms for next level analysis.

Term based Polarity Identification

The result vectors are processed to identify their polarity levels. Polarity identification is performed using the polarity repository, SentiWordNet 3.0 [25]. The SentiWordNet being a human annotated data repository, has very high reliability levels. Polarity identification process has a major issue associated with it. Not all elements can be directly provided a polarity level. Under certain instances, certain terms exhibit positive polarity, while the same terms can exhibit negative polarity under other instances. Hence every term in SentiWordNet is associated with a positive and a negative polarity level. The polarity level of a term in the result vectors is selected based on the polarity of the input query. Positively or negatively polarized queries results in their corresponding polarity value to be retrieved. If the query exhibits neutral sentiment, both the polarity values are considered and the final polarity is identified using eq. 4

$$Polarity_t = Polarity_{(pos,t)} - Polarity_{(neg,t)} \tag{4}$$

Where Polarity(pos,t) refers to the positive polarity associated with the term t and Polarity(neg,t) refers to the negative polarity associated with the term t.

Significance based Weighted Sentiment Identification

Polarity levels and the significance of the term in the context are used to identify the sentiment associated with the result vectors. Sentiment associated with a result vector v is given by the cumulative sum of the product of its polarity and its significance (TF-IDF).

$$Sentiment_v = Polarity_{(t)} * TFIDF_{(t)} \forall v = 1, 2 \ldots n \qquad (5)$$

Hence the polarity of a more significant term provides higher impact on the sentiment of the vector compared to the polarity of a lower significant term. The identified sentiment value directly corresponds to the rank of the term in the retrieved results.

2.2.3 Profile based Result Ranking

The sentiment intensity based results ranks the vectors in terms of their significance. However, their association with the user's interests is still not incorporated into the analysis process. The profile based result ranking module correlates the retrieved results with the user's profile to identify results with high correlation levels. This is performed in three phases. Vectors with terms contained in the user's profile are shortlisted as Level 1 selections. Vectors in Level 1 are again shortlisted based on the polarity of the input query and are categorized as Level 2 selections. The final phase sorts vectors in each of the levels based on their vector sentiment obtained from eq. 5. Level 2 vectors occupy the top slot, followed by Level 1 vectors and then by the other input vectors. The ordered results are presented to the user.

2.2.4 Profile Update and Reinforcement

User selections are monitored and selection of a result results in the increase in frequency of the term in the user's profile. This relevance feedback updates the user's profile, hence building it such that the learning reinforces the existing rules and creates new rules for effective identification of the user's interests [26]. Every search made by the user impacts the profile, leading to highly customized results appropriate to the user's interests.

3 Results and discussion

The proposed architecture was implemented in PySpark and the results were obtained using the STS Gold Sentiment Corpus. The ROC plot in figure 2 shows the prediction level of the proposed sentiment identification module. It could be observed that, though the true positive levels were initially low, the levels increase to a maximum of 0.87, exhibiting effective prediction levels. Low to moderate false positive levels can be observed indicating the robustness of the proposed model.

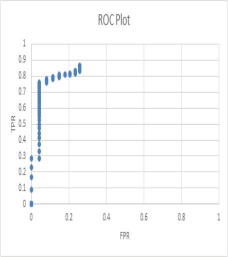

Fig. 2: ROC Plot

The PR plot shown in figure 3 exhibits the retrieval levels of the proposed model. A high precision and a high recall level indicates effective prediction. It could be observed that the proposed prediction model exhibits a very high precision level of 1 and a high recall level of 0.9. This exhibits retrieval of highly relevant results by the prediction model.

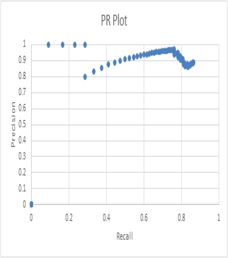

Fig. 3: PR Plot

Scalability levels of the proposed model is shown in figure 4. Queries beginning from low complexity to high complexity are provided to the model and the time taken for processing was recorded. Query with low complexity indicates a direct query term. Vagueness in terms are increased, with increase in complexity levels and query terms are replaced with multiple term text. It could be observed that increase in complexity leads to increase in processing times, due to the higher number of results retrieved.

A comparison between the emoticon based sentiment analysis model and the proposed reinforcement based information retrieval model has been performed and the resultant metrics are shown in figure 5. It could be observed that the proposed model exhibits high accuracy levels, F-Measure levels and TNR levels indicating efficiency in the prediction process. Very low false negative rates also indicate lowered error levels in the prediction process.

© 2017 NSP

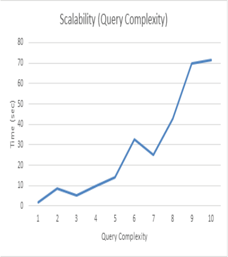

Fig. 4: Scalability Analysis

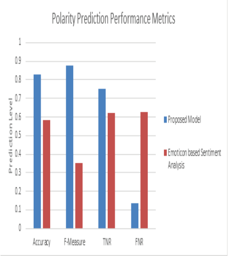

Fig. 5: Performance Comparison

4 Conclusion

This paper presents an effective information retrieval model that operates based on user's feedback in-order to provide highly correlated results to the user. Every query and the subsequent selection of result, builds the user's profile. This leads to an experience, tailored specifically for the user. Occasionally the user might also perform searches out-of-context, therefore result elimination is avoided. All the results are presented to the user in the order of relevance. Top results are of high relevance, while results of low relevance can also be obtained. The limitations of this model are; it can operate only on structured data and only on content written in English. Future extensions of the proposed model can include operating on semi-structured or unstructured data. This can also be extended to multimedia content rather than plain text. Cross lingual retrieval rules can also be incorporated into the model to enable CLIR.

References

[1] S. Marrara, G. Pasi, and M. Viviani, Aggregation operators in Information Retrieval. Fuzzy Sets and Systems.,2016

[2] A. Bigot, C. Chrisment, T. Dkaki, G. Hubert, and J. Mothe, Fusing different information retrieval systems according to query-topics: a study based on correlation in information retrieval systems and TREC topics. Information Retrieval, 14(6), p.617. 2011

[3] Hussein, D.M.E.D.M. A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences. 2016

[4] B. Jiang, J.Yang, , Z. Lv, , K. Tian, Q. Meng, and Y. Yan, Y. Internet cross-media retrieval based on deep learning. Journal of Visual Communication and Image Representation. 2017

[5] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in:International Conference on Multimedia, 2010, pp. 251260.

[6] K. Chaudhuri, S.M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June, 2009, pp. 129136.

[7] M.R. Bouadjenek, H. Hacid, and M. Bouzeghoub,. Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms. Information Systems, 56, 2016 pp.1-18.

[8] C.Lioma, R.Blanco, M.-F.Moens, A logical inference approach to query expansion with social tags, in: ICTIR, 2009.

[9] S.Jin, H. Lin, S. Su, Query expansion based on folksonomy tagco-occurrence analysis, in: 2009 IEEE International Conference on Granular Computing, 2009.

[10] A.Mantrach, J.-M.Renders, A general framework for people retrieval in social media with multiple roles, in :R.Baeza-Yates, A. de Vries, H.Zaragoza, B.Cambazoglu, V.Murdock, R.Lempel, F. Silvestri(Eds.), Advances in Information Retrieval, Lecture Notesin Computer Science, vol.7224, Springer, Berlin, Heidelberg, 2012, pp. 512516.

[11] Y.Lin, H.Lin, S.Jin, Z.Ye, Social an notation in query expansion: a machine learning approach, in: Proceedings of the 34th International ACMSIGIR Conference on Research and Development in Information Retrieval, SIGIR'11, ACM, NewYork, NY, USA, 2011, pp.405414.

[12] M. Madankar, M.B. Chandak, and N. Chavhan. Information Retrieval System and Machine Translation: A Review. Procedia Computer Science, 78, 2016 pp.845-850.

[13] M.S.Madankar. A Review on Information Retrieval in Indian Multilingual Languages. International Journal of Advanced Research in Computer Science and Software Engineering. Vol- 5 Issue 3; March 2015.

[14] 3. Monika Sharma, Dr. Sudha Morwal. Refinement of Search Results of the Google using Cross Lingual Reference Technique and GPS. International Journal of Emerging Research in Management, Technology, Vol-4, Issue-4, April 2015.

[15] 4. Savita C. Mayanale, S. S. Pawar. Survey on Indian CLIR and MT systems in Marathi Language. International Journal of Computer Applications Technology and Research. Vol-4, Issue 7, 579 - 583, 2015.

[16] L. Guezouli, and H. Essafi. CAS-based information retrieval in semi-structured documents: CASISS model. Journal of Innovation in Digital Ecosystems, 3(2), 2016 pp.155-162.

[17] R.D. Burke, K.J. Hammond, E. Cooper, Knowledge-based information retrieval from semi-structured text, in: In AAAI Workshop on Internet-based Information Systems, AAAI, 1995, pp. 1519.

[18] Nirmal, V. Jude, and DI George Amalarethinam. "Emoticon based Sentiment Analysis using Parallel Analytics on Hadoop." Indian Journal of Science and Technology 9.33, 2016.

© 2017 NSP
Natural Sciences Publishing Cor.

[19] F. Zhou, J.R. Jiao, X.J. Yang, and B. Lei. Augmenting Feature Model through Customer Preference Mining by Hybrid Sentiment Analysis. Expert Systems with Applications. 2017

[20] N. Hariri, C. Castro-Herrera, M. Mirakhorli, J. Cleland-Huang, B. Mobasher, Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings. IEEE Trans. Softw. Eng., Vol-39, pp. 1736-1752., 2013.

[21] J. Bollen, H. Mao, X. Zeng. Twitter mood predicts the stock market. Journal of Computational Science, Vol-2, pp. 1-8, 2011

[22] P.J. Stone, D.C. Dunphry, M.S.Smith, D.M. Ogilvie. Cambridge, MA: MIT Press. 1966

[23] O. Alhabashneh, R. Iqbal, F. Doctor, and A. James. Fuzzy rule based profiling approach for enterprise information seeking and retrieval. Information Sciences, 394, 2017 pp.18-37.

[24] R. Baeza-Yates, and B. Ribeiro-Neto. Modern information retrieval (Vol. 463). New York: ACM press. 1999

[25] S. Baccianella, A. Esuli, and F. Sebastiani, May. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC 2010, Vol. 10, pp. 2200-2204.

[26] L. Zhang, Y. Zhang, Q. Xing, Filtering semi-structured documents based on faceted feedback, in: W.-Y. Ma, J.-Y. Nie,R.A. Baeza-Yates, T.-S. Chua, W.B. Croft (Eds.), SIGIR, ACM, 2011, pp. 645654.

S. Subitha has received her Master of Philosophy (M.Phil) in Computer Science from Periyar University, India in the year 2008 and also her Post Graduate Degree (MCA) from Bharathidasan University , India in the year 1997. Presently she is a research scholar of Anna University Chennai. She has published 8 papers in national and international conferences and 3 papers in international Journals. She is a keen researcher in web data mining techniques.

S. Sujatha is a Doctorate in Computer Science and having twenty years of experience in teaching with good knowledge in the area of Computer science and Information Technology, currently working as Associate Professor, Bharathidasan Institute of Technology, Anna University, Tirchirappalli, India. She has published 20 research papers in reputed journals, international and national conferences. She has received .Active Researcher Award, Anna University, Chennai in the year 2013. She is a recognized research supervisor in Anna University, Chennai. Her areas of interest include Distributed Computing, Web services, Data Mining, Cloud Computing and its applications

© 2017 NSP
Natural Sciences Publishing Cor.