Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (3): 485-490, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.485.490

Intelligent Query Refine and Expansion Model for the Retrieval of E-learning Resources

[1]T. Chellatamilan and [2]R.M. Suresh
[1]Research Scholar, Department of Computer Science and Engineering,
Arunai Engineering College, Tiruvannamalai, India
[2]Principal, Jerusalem Engineering College, Chennai-32, India

Abstract: Today Internet Search Engines, employing intelligent and innovative programming skills, do almost exhaustive search of data contained in websites through number crunching operations at enormous speed but as the search results usually run into hundreds of pages, it is not therefore always possible for an average user to make effective use of such huge collections of data produced as output. While input queries to the Search Engines are usually in the form bag of words, most Information Retrieval systems of Search Engines use models based on statistics of word counts for identifying and ranking the documents in the order of their relevance. In order to make further progress and achieve refinements, researchers have extensively studied the application of Artificial Intelligence (AI), concept of Intelligent Agent (IA) to Information Retrieval Tasks. In this study, the query reformulation has been performed based on the statistical probabilistic language modeling technique. The comparative study on the precision recall before and after query expansion been studied and the plot has been plotted.

Key words: Mobile agents, e-learning, artificial intelligence, query, retrieval

INTRODUCTION

Relevance of information is a critical factor in knowledge assimilation and presentation (Zhai, 2008). With the current available repositories of knowledge in the forms of documents emphasis towards identifying, gathering and analyzing and presenting the content are key factors. A typical eLearning scenario requires all the previously mentioned factors to be optimized. The current scenario of eLearning system with abundant digital information, the learner finds it difficult to fetch relevant documents which associate more towards the requirement. The need of the hour is to have an intelligent self learning system which can present the user with highly relevant feedback based on the user query (Heie and Edward, 2012).

The field of Information Retrieval (IR) is studied intensely as huge amount of data is made available in electronic format in Internet and various other sources. Statistical Language models have been successfully applied in various IR tasks. The recent studies show that the use of statistical language models yield superior empirical performance and these models can also be made adaptable for carrying out training and learning functions. Many variations of the language models are developed and applied in multiple information retrieval tasks. The tasks include cross-lingual retrieval, distributed IR, expert finding, web search, topic tracking and subtopic retrieval.

In this research work, language model approaches are studied for information retrieval tasks in the field of e-Learning. The goal of an Information Retrieval (IR) system is to rank documents optimally given a query so that relevant documents would be ranked above the non-relevant documents. In order to achieve this goal, the system must be able to score documents so that a relevant document would have a higher score than a non-relevant document. Clearly the retrieval accuracy of an IR system is directly determined by the query of the scoring function adopted. Seeking optimal function (Retrieval Function) has always been a major challenge in information retrieval.

In the statistical model, the joint probability distribution P (D, Q) is analyzed and computed in terms of P (Q|D), where the random variables D and Q represent "Document" and "Query", respectively (Zhai and Lafferty, 2002). This leads to a situation where the probability of generating the query from the document associated with the relevance of the document. A document which is more likely to generate the query is assumed to be more relevant.

In an alternate statistical model, the joint probability distribution P (D, Q) is analyzed and computed in terms of

Corresponding Author: T. Chellatamilan, Research Scholar, Department of Computer Science and Engineering, Arunai Engineering College, Tiruvannamalai, India

P (D|Q). This leads to a situation of estimating the probability of classifying a relevant document where user's information need is given. The choice of this model has implications for training the parameters.

For the two types statistical models discussed above, several improvements to the existing scoring functions are attempted and the results are presented.

As Information sources in the e-Learning environment are usually widely distributed, the adoption of statistical models based on mobile intelligent agent approach is also studied.

While most of the information sources are of textual data type for e-Learning applications, information is also provided through highly organized and structured data base formats. In view of this, the Information Retrieval system is also studied as Multiple Attribute and Multi Criteria Decision making process in order to explore the usefulness of applying model based IA tools for identifying and ranking the relevant documents. In this respect, the adoption of IR system based on a Boolean key word model is also studied (Li *et al*., 2012).

The Basic IR Systems uses the TF-IDF, TF-IDF Weighting, Vector space Model, BM25 and etc. for ranking the documents optimally for the given query (He and Ounis, 2004). The Statistical Language Modeling Technique also helps for answer retrieval and answer classification from fixed corpus (Heie and Edward, 2012). When the Queries are expanded, language modeling gives better results when using probabilistic dictionary or the relevance feedback (Larkey and Connell, 2005). Different smoothing technique has been incorporated into the IR System to improve the score in ranking the documents (Mei *et al*., 2007).

The query and its results are analyzed and its semantic distance between them are evaluated to categorize the query which inherently shows the similarity between them (Li *et al*., 2012). The Agent based information processing system increases information availability and uncertainty through information scanning, filtering, interpretation and alerting (Mark *et al*., 2011). The Strength of relations between two linked documents can be obtained by a relational clustering algorithms based on probabilistic graph r epresentations with k-means and expectation maximization techniques (Fersini *et al*., 2010). The Learning objects/contents in an e-learning systems are being classified or ranked based on relevance ranking metrics like topic, personal and situational relevance (Ocoha and Duval, 2008). The machine learning and data mining approach has been used to form a relationship hierarchy of all the concepts represented by the learning material. The classified and co-related learning materials are recommended to the peer learners then (Hsieh and Wang 2010). To correct the misspelled query, the linguistic information is required to be incorporated into the IR system (Vilares *et al*., 2011) Term weighting scheme improves the conventional TF*IDF and language models through evidential term weights in the collection statistics (Song and Myaeng, 2012). Rather than using Term weights and the relevance feedback the wordnet light ontology has also been used for the query expansion (Dragoni *et al*., 2012).

The research work aimed at describing the techniques for improving the relevance of the search results, speed at which the content is presented to the user and suggestion on possible accurate queries that can be executed in the future.

SYSTEM ARCHITECTURE

In relevance to the overall goal of improving the search results for the user query, the system is defined as follows. The leaner observes the requirement to initiate a search for a document and submits a query to the learner interface. The learner interface primarily acts as the source point of the agent creation and distribution before it can be distributed to the network. The user relies on the agent to collect the information to assimilate the information returned and to be presented in user relevant format. The agent distributed over the network is accepted by a retrieval agent which authenticates the request and the source of the agent. On successful authentication, the query submitted by the user becomes authorized to be submitted to the IR (Information Retrieval) system for processing. The IR system is a combination of Query likelihood estimation and language modeling engine with the help of parameterization using probability distribution models like Poisson or multinomial.

The query is submitted to the Document term likelihood Analytical Engine (DLAE). The DLAE is an independent engine which periodically classifies and attributes the different documents within a repository. It statistically maintains the probability occurrence of different terms within document and across the document collection in the document corpus. Every document corpus will have an associated DLAE engine thereby ensuring that the statistical information available on the probability of available terms are updated whenever there is a new document added or an available document is deleted from corpus or repository.

The DLAE in Fig. 1 returns the document set that matched the query to the IR System. Within the IR system, the ranking engine understands and learns the ranking of the documents retrieved in relevance to the query submitted. The result set of documents are

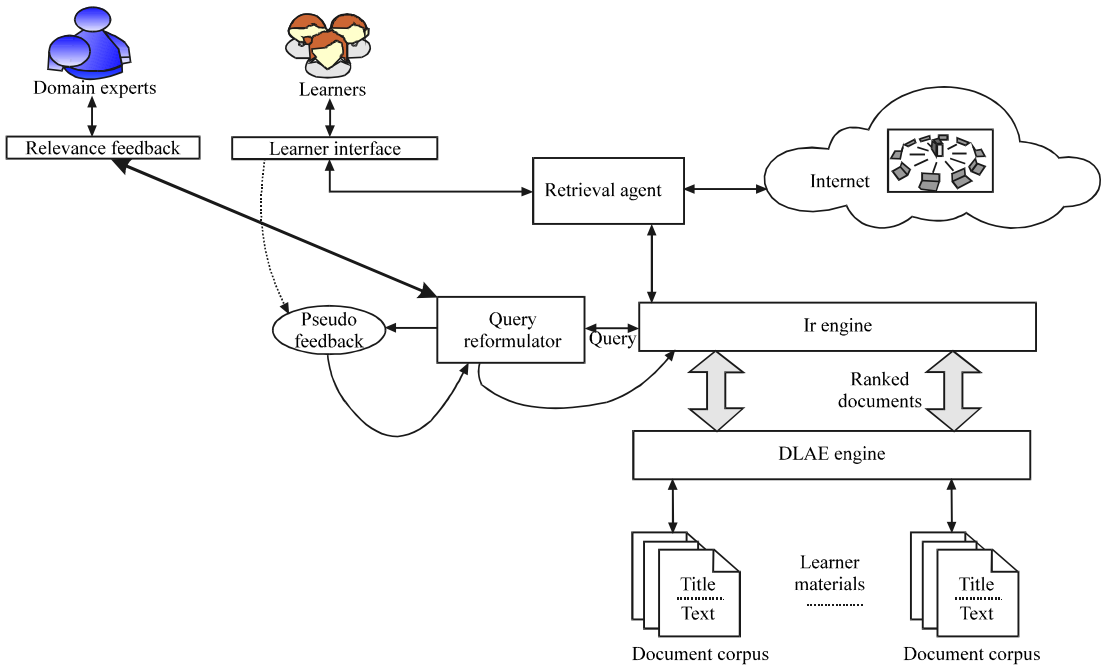

Fig. 1: Block diagram

submitted to relevance comparator which compares the results to the relevant feedback from the domain experts. If the retrieved result set does not match at least 50% relevance to the relevance feedback results, the system submits the query back to the query expansion and re-formulator to identify a more suggestive query phrase. This newly generated query is submitted to IR interface to retrieve new results which includes the re-ranking of the documents. The final result set is submitted to the learner interface when the relevance is above 80% in comparison to the relevance feedback. This result set is presented to the user along with the query expansion suggestions. The user feedback on the result set or initiates a feedback based on the Query Expansion suggestion provided. This pseudo feedback from the user is used by the query re-formulator engine to reconstruct the query to make a more relevant and precise search as shown in the Fig. 1. The expanded query is submitted to the IR engine to go through the same process thereby allowing the ranking engine to re-rank the documents retrieved and submit to the learner interface with a result set which is more relevant and precise to the requirement.

SYSTEM COMPONENTS

In a pedagogical environment, the learner submits a query to search for documents that contain the query phrase. The challenge is to present the user with the most relevant documents that matched to the submitted query phrase and suggestive alternates for reconstructing the query in order to retrieve suitable and better alternatives. The learner becomes an active participant in evolving the system to learn relevance with respect to the search phrase and available documents which represents the training phase of the machine learning process.

Learner interface: The learner submits the query to the learning interface which initiates the distribution of the mobile agents with the query construct to the retrieval agents. The learner interface is a bridge which alternates between agent management and presentation of the result set to the user.

The Learner Interface extracts the results (from the IR system) collected through the retrieval agent to differentiate between the result set that matched the current query construct and the suggestive query expansions provided by the IR system. Interfacing the core Intelligence engine and the learner, the learner interface not only acts as a bridge but also as a learning system identifying the destination of the agents for future distribution.

Learning Objects Repository (LOR): The Learning objects repository is a collection of documents on which the user query will be executed in order to identify documents that contain the query construct. The repository can contain documents in different knowledge areas which may or may not be relevant to the current

query phrase. The overall system can contain n document corpus each with an associated set of documents on which the query could be executed. The LOR will be accessed by the IR system through the DLAE (Document term Likelihood Analytical Engine).

Information retrieval system: The user submitted query is passed by the hosting Learner interface through the Retrieval agent to the Information Retrieval (IR) system. The IR system is responsible to return the most likely occurrences of documents wherein the search phrase best fits. Before proceedings further to understand the working of the IR system, it is important the Language Model is understood. It comprises of:

- Leaner interface
- IR systems and language modeling
- DLAE (Document term Likelihood Analytical Engine)
- Learning object repository
- Query Re-formulator and expansion
  - Learning object repository
  - Information retrieval and machine learning
  - Query generation Language model
  - Probabilistic distribution
  - Query expansion or reformulation through relevance feedback model
  - Experimental set up and result
  - Conclusion

Language modeling: Language modeling is the probability distribution defined over a particular vocabulary. In our scenario we focus on the Unigram Language Model (ULM) where one predict the likelihood occurrences of every word independent of the other. After predicting the individuality of words the ULM also assigns the probability phrases by multiplying the probabilities of the individual words contained in the phrase. The ULM consists of two main steps:

- **Estimation:** Estimating probabilities of each word
- **Prediction:** Assigns probability to span of text or phrase

The ULM estimations requires to first count every individual word is considered as count and then the total term is calculated every term is assigned a probability by valuing the number of occurrences of each terms against the total number of term occurrences which becomes the probability of each term.

Document ranking principle: The ranking engine uses the ULM to estimate the score of each document relevant to the query based on the probability that it close to atisfy the query term:

$$\text{Score}(Q,D) = \prod_{w \in Q} P(w \mid \theta_d)$$

The ranking engine refines the result set by categorizing the document list in the descending ranking order. Non occurrences of a query term (qi) in the document can result in cumulative score of zero probability in the ULM. To avoid zero probability error, we resort to smoothing probability estimation. The goal of smoothing is to decrease the probability of observed outcome and to increase the probability of unobserved outcome. The idea is to allocate some probability to the unobserved terms and reduce probability to those terms appearing in the document collection (Collection language modeling). Injecting the non occurrence term into every document at least once, ensures that the probability out-come in the language modeling is never zero. An alternative approach is to use the linear interpolation smoothing where redundant injunction of non occurrence term can be avoided. Linear interpolation smoothing calculates the probability of occurrences of a term on an individual document along with the occurrence of the term over the entire document corpus or collection.

P_t = TF(t, d)/N
P_t = Probability of the Term t in doc ‘d’
TF(t, d) = Term Frequency of the term in d
N = Total number of terms in ‘d’

The linear interpolating smoothing allows to rank the document based on the score assigned towards it in the query likelihood language model. The document score is the probability that it generated the query by considering all terms occurrences with in the query submitted.

$$\text{Score}(Q, D) = \alpha P(w|\theta_d)+(1-\alpha)P(w|\theta_c)$$

The query likelihood relevant model ensures the term which occurs less frequent in the entire collection have a highest contribution towards the document scores.

The DLAE stages over every document repository and creates the probability distribution statistics for every document with in the document collection. Based on the query submitted, the DLAE calculates the query like hood score to construct the result set of documents to be retrieved.

To refine the computation of the likelihood on a document language model for an occurrence of the query several distribution techniques can be used.

Binomial/multi Bernoulli Probability distribution-This model estimates the parameter which denotes the probability of the query term considering its presence or absence. Multinomial probability distribution also considers the frequency or a count of the query terms presence in the language model which estimates the parameter.

Poisson model: There are few problems encountered, the multinomial distribution fails to address model term absence sum-to-one overall terms.

Query with poisson: As described already, on every document in the document corpus, the language model created provides the statistics of the occurrence and individual terms.

In comparison to arrive to the query substitution of the user, the rate of arrival of every query terms is calculated to arrive at the probability value for the document with respect to the query P(q|d):

$$P(d|\theta) = \prod_{i=1}^{n} \frac{(\lambda \mid d|)^{c(wi,d)} e^{-\lambda|d|}}{c(w_i,d)!}$$

The fundamental smoothing approach is followed to ensure the score of a document is further refined, considering the background collection model. The refined set of documents retrieved by the DLAE based on the score is compared by the IR Engine to the relevant feedback from the word class dataset for query relevance. If the resulting set of documents does not match at least 70% of the relevant feedback the query formulator expands the query to be re-submitted to the IR Engine.

The reformulation of the query is done through term re-weightage. It identifies and uses the terms with higher weightage in the relevance feedback. When the IR identifies final result set as at least 80% match to the relevance feedback it submits the result to the retrieval agents which in turn propagates it to the user interface.

The resubmitted query is processed by IR Engine and the documents are re-ranked based on the new query construct. The result set includes documents and iterative terms identified by the query re-formulator. The user is presented with possible query expansions that were considered when retrieving the documents. A search initiation by the user on any one of these query expansion phrase possibilities is feedback to query re-formulator as a pseudo feedback.

EXPERIMENT

To demonstrate the query refinement and expansion model using Poisson distribution, we have taken the following experiment. The experiment takes the dataset available for TREC 3 on the research journal paper as the document corpus. The documents are categorized under different classification. Under every classification there is a set of 100 documents which will be used for training the query engine's preprocessing to generate LM.

A python program has been used to generate the language model for each document available to the document corpus. Python has been chosen as the programming language due to its inherent capabilities for information retrieval search. As open source software it allows the user to extend the limitations of the program for their requirement and future enhancement.

The resulting LM generated based on the probability calculated for the individual term is shown in the Table 1.

Testing phase: A refined set of 30 documents in each category is identified to execute the experiments and compare query results to that of TREC 3 evolution bench mark.

A set of 64 queries have been identified as published by TREC 3 to be executed against the test documents for corpus analysis. The initial phase of reevaluating the ranking of documents against the test query is performed and mentioned in Table 1. Table 1 is without IL query expansion process. From the result reference we see, that the number of reference documents refined materials less than 70% of the expected results of standardized of TREC. To follow refer the results set to match a retrieval result at least to 70% of the evaluated benchmark, the

Table 1: Probability calculation for the individual term

Doc. Name	No. of Doc.	No. of Queries	Size in MB
CACM	3204	64	1.5
CISI	2704	34	1
CRAN	1204	23	1
MED	999	43	1

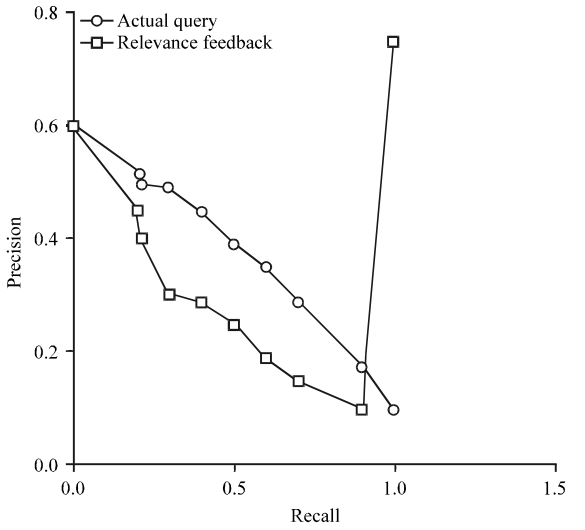

Fig. 2: Refined result set and the comparative chart to the evaluation bench mark

query is resubmitted with the following possible query expansion. The refined result set and the comparative chart to the evaluation bench mark given in Fig. 2. This process is repeated until the user is able to reach 70% of the benchmark.

CONCLUSION

The experiments allow us to compare the probability by using different query expansion and probability statistical techniques for continuous search based on the query input. Our analysis shows that the probability calculator based on the Poisson distribution provides more accurate and relevant search results compared to other technique. In a pedagogical environment the learner not only associates himself with the results retrieved but also to the other possible queries which can be used in future that can retrieve the document faster and relevant. The scope of the experiment and work though limited to the query retrieval technique does open answers to possibilities where query reformulation and continuous learning can be enhanced. All the experiments were based on the SMART collection Dataset (ftp:\\ftp.cs.cornell.edu\pub\smart).

REFERENCES

Dragoni, M., C.C. Pereira and A.G.B. Tettamanzi, 2012. A conceptual representation of documents and queries for information retrieval systems by using light ontologies. Expert Syst. Appl., 39: 10376-10388.

Fersini, E., E. Messina and F. Archetti, 2010. A probabilistic relational approach for web document clustering. Inform. Proc. Manage., 46: 117-130.

He, B. and I. Ounis, 2004. A query-based pre-retrieval model selection approach to information retrieval. Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval, April 26-28, 2004, University of Avignon, France, pp: 706-719.

Heie, M.H. and W.D. Edward, 2012. Questions and answering System using statistical language modeling. Comput. Speech Lang., 26: 193-209.

Hsieh, T.C. and T.I. Wang, 2010. A mining-based approach on discovering courses pattern for constructing suitable learning path. Expert Syst. Appl., 37: 4156-4167.

Larkey, L.S. and M.E. Connell, 2005. Structured queries, language modeling and relevance modeling in cross lingual information Retrieval. Inform. Process. Manage., 41: 457-473.

Li, L. and L. Zhog, G. Xu and M. Kitsuregawa, 2012. A feature free search query classification approach using semantic distance. Expert Syst. Appl., 39: 10739-10748.

Mark, X.U., V. Org, Y. Duan and B. Mathews, 2011. Intelligent Agent Systems for executive information scanning, filtering and interpretation perceptions and challenges. Inform. Process. Manage., 47: 186-201.

Mei, Q., H. Fang and C. Zhai, 2007. A study of poisson query generation model for information retrieval. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 23-27, 2007, ACM, New York, USA.

Ocoha, X. and E. Duval, 2008. Relevance ranking metrics for learning objects. IEEE Trans. Learn. Technol., 1: 34-48.

Song, S.K. and S.H. Myaeng, 2012. A novel term weighting scheme based on discrimination power obtained from past retrieval results. Inform. Process. Manage., 48: 919-930.

Vilares, J., M. Vilares and J. Otero, 2011. Managing misspelled queries in IR applications. Inform. Process. Mange., 47: 263-286.

Zhai, C. and J. Lafferty, 2002. Two-stage language models for information retrieval. Proceedings of the 25th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, pp: 49-56.

Zhai, C., 2008. Statistical language models for information retrieval: A critical review. Found. Trends Inform. Retrieval, 2: 137-213.