International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(3), pp.208-214
DOI: http://dx.doi.org/10.21172/1.83.031
e-ISSN:2278-621X

COMPARATIVE STUDY FOR RESEARCH IN INFORMATION RETRIEVAL

Shiwani Gupta[1]

Abstract— Relevant document search through relevant query at the interface is an essential aspect for research in the field of Information Retrieval (IR). Thus the author has tried to explore and compare possible avenues for research in the area considering works of peers in the field. The author after studying enormously has identified following areas : Relevance Feedback, Retrieval Optimization and Feedback Models where improvements can be done. Further collation of datasets and evaluation metrics has been carried out to ease out experimentation in the area. The author assumes that researchers in the field would benefit from the composite work presented here and could focus on Optimizations in IR effectiveness.

Keywords—Information Retrieval; Relevance Feedback; Optimization

I. OBJECTIVE

The aim of this paper is to focus on the various knowledge required for research in Information Retrieval, a field where optimization could be w.r.t. to time of retrieval or effectiveness in retrieval. Though effectivenss is subjective, a lot of work has been done to improve retrieval effectiveness by studying relevance feedback models, of which Implicit feedback has been worked upon the most. Hence study required for various Implicit feedback models. Another research could be to access various optimization techniques applicable for IR. Different researchers have worked upon different datasets for their experiments, so which to use when and there are no. of evaluation metrics applicable in the field, so which is the one that judges the best, etc.

II. INTRODUCTION TO INFORMATION RETRIEVAL

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). It is becoming the dominant form of information access, overtaking traditional database style searching. Examples include library catalogs, patents, research articles, email programs, web search etc.

Standard terms in IR:

An **inverted index** or inverted file maps an index from term to parts of document they occur. Construction steps are:

- Collect documents to be indexed
- Tokenize text (Determine dictionary / **vocabulary** of terms)
- Drop common terms (**stop words**)
- **Stemming**
- Do linguistic preprocessing of tokens
- Index documents that each term occurs in (**Posting**)

[1] *A.P. CMPN TCET Mumbai, India*

Retrieval effectiveness could be captured through Relevance feedback that involves an iterative process in which users first specify which documents are relevant to them. These specified documents are used by the system to retrieve more or similar documents. The process is then repeated. Relevance feedback may be **Explicit** where users may be constrained and reluctant to provide feedback or **Implicit** where user's click-through record provides feedback but it is still under exploration that what interactions should be taken into account and how good they are as relevance feedback or it may be **Pseudo** which assumes top ranked documents as relevant but doesn't guarantee it. When the user's interactions are removed from this iterative process, this kind of relevance feedback is called PRF or blind relevance feedback.

III. LITERATURE SURVEY

In the field of Information Retrieval, researchers have worked upon few challenges as Relevance modeling and Irrelevance modeling; when relevant documents are less. In [1], authors have come up with approximate true relevance model, developed its framework, theory and Distribution Separation method to separate mixed probability distribution of relevant and irrelevant documents through relevance feedback which is a post query process by building refined query model; to improve the retrieval effectiveness and/or stability of query model estimation in the context of relevance feedback.

In [2], authors have evaluated uncertainty and risk in Information Retrieval.

The objective of query expansion is to extend a first, unsuccessful query with different words that best catch the user's purpose, or that creates a relevant query that will likely retrieve more significant documents. Explicit relevance feedback is repeated until user is satisfied with outcomes. Since user evaluates importance of document, it puts enormous weight on user. Thus, an alternate technique Pseudo relevance feedback can be used which is implicit and assumes that top results have higher accuracy and features. [3]

The simulation-based evaluation methodology measures how well the models learn what information is relevant across different documents and how well they create effective search queries through document representations which include document title and query-biased summary of the document; a list of top-ranking sentences (TRS) extracted from the top documents retrieved, scored in relation to the query; a sentence in the document summary, and each summary sentence in the context it occurs in the document (i.e., with the preceding and following sentence). These representations allow searchers to more deeply explore the retrieved information and can combine to form an interactive relevance path at the search interface. A relevance path is traversed if searchers travel between different representations of the same document. The paths provide searchers with progressively more information from the best documents to help them choose new query terms and select what new information to view. [4]

In [5], authors have proposed a query generation method for searching scientific datasets. The conventional text-based pseudo relevance feedback (PRF) methods are extended by using spatial and temporal information. The proposed method scores the spatial and temporal distances according to the Bhattacharyya distance among datasets and uses this information to rank the search results. The search interface designed retrieves a ranked list of documents and presents these with associated titles and snippets, visualizes the query and the retrieves documents as a mixture of topics. The sections of the query are shown with associated topic classes. Secondly, the system lets a user select a particular topic of his choice for feedback. As a result of this topic-based feedback, the documents are reranked. Thirdly, the interface provides easy access to document content by letting a user follow hyperlinks from one part of a document to another on the same topic.

IV. COMPARISON RESULTS

Table I. RESEARCH TRENDS IN RELEVANCE FEEDBACK FOR INFORMATION RETRIEVAL

Problem	Methodology	Future Scope
Mixed probability distribution of relevant and irrelevant documents even in relevant feedback set.	**Distribution Separation Method through explicit relevance / irrelevance feedback. Automatic irrelevance feedback - simulated and explicit. Document reranking to rank irrelevant documents low.**	**Graded relevance. Applying DSM to online tweets filtering and IR tasks.**
Quality of seed irrelevant distribution is also not	**Penalized weighted precision of irrelevance to measure quality of seed irrelevant document.**	

guaranteed. [1]	**DSM Regularization framework where probabilities of relevant terms are made more dominant, which enforces estimated relevance distribution to be close to reference relevance distribution and mixture distribution controls possible estimation of relevance distribution.**	
Retrieval and estimation effectiveness and stability tradeoff. Risk management. Overall estimation quality and risk (Irrelevant document removal). Detect seed irrelevant documents. [2]	**Distribution Separation Method Pseudo relevance feedback Document smoothing method. Bias – Variance analysis for expanded query model. Outlier document and term detection to extract irrelevant documents.**	**Apply Bias variance analysis to personalization. Vary relevance estimation for different users in different contexts during different times.**
Microblog are Short document Informal language Lack of contextual clues Vocabulary mismatch. Real time nature. [3]	**Pseudo relevance feedback model employing LDA. Topic based Query expansion. Language model. Temporal and lexical feedback**	**Incorporate temporal evidence in TBPRF.**
Searchers have to explicitly mark documents as relevant Traditional measures as document reading time and scrolling can be unreliable and context dependent. [4]	**Implicit feedback model: Query modification is based on inferences made from searcher interaction. Information in form of document representation not full document at retrieval time. Learn relevance and create more effective search queries**	**Development of models of behavior to represent different situations, searchers, and searching styles**
Scientific data contains relatively little text information. [5]	**Query generation method using spatial, temporal, and text information based on pseudo relevance feedback.**	**KL divergence and Hellinger distance can be applied for measuring distance b/w datasets. Can add ontological and citational correlation**
Multitopical characteristics of retrieved documents in patent prior art search. [6]	**Sentence based query expansion SBQE Topical relevance model TRLM (Topic visualization, topic based feedback based on document and query segmentation by TextTiling method, topic navigation) Interactive search interface Pseudo Relevance Feedback**	**Applying a round-robin technique weighted by the similarities between documents and query segments Hierarchic topic modeling**

Table II. DATA SETS FOR INFORMATION RETRIEVAL

Data sets	Specialty
TREC WSJ8792 [1,2]	**173252 documents**
TREC AP8889 [1,2]	**164597 documents**
TREC ROBUST2004 [1,2]	**528115 documents**
TREC WT10G [1,2]	**1692096 documents**
TREC SJMN [4]	**Topics 101-150**
TREC 2011, 2012 [3]	**16 million tweets published in various languages over a period of 2 weeks**
World Data System [5]	**Scientific datasets from more than 100 standalone datacenters**
Pangaea [5]	**Environmental science datasets**
TREC Vol 4, 5 [6]	**528542 American news article**
CLEF-IP 2010 data [6]	**patent prior art search data set**
CISI [11]	**1640 documents**
CACM [11,12]	**Small (3204 documents)**
Reuters-RCV1 [10]	**800,000 documents**
NPL [11]	**Large**
Yahoo query click log [12]	**Yahoo! training set contained almost 40 million clicks performed during the first 20 days of July 2010 (excluded), whereas the evaluation set contained the remaining 26 million clicks issued until the end of the same month**
AOL query click log [12]	**30 million clicks issued by 650,000 users, recorded in a times- pan going from March to May 2006**

Table III. EVALUATION METRICS FOR INFORMATION RETRIEVAL

Evaluation Metrics	Inference
Mean Average Precision [1,3,6]	mean value of average precision over all queries and reflects the overall retrieval effectiveness
Normalized Discounted Cumulative Gain @ N [3,7,9, 12]	Compares and votes results to rank search engines N=1, 3, 10

Precision @ N [3,6,13,14]	N=10, 30
Recall @N [5,6]	Can only handle cases with binary judgment
Pearson correlation [9]	b/w human and automatic evaluation
Non-parametric correlation coefficients Spearman's rho, Kendall's tau-b [4]	Measure rate of learning
Patent Retrieval Evaluation Score [6]	Per topic feedback averaged over topics
Bhattacharya Distance [5]	one of the standard metric used to measure the distance between two probability distributions
$Bias^2$ + variance [2]	Robust metric
F-measure [6]	Weighted harmonic mean of precision and recall
PRES [6]	Recall oriented metric overcomes excessive precision bias of MAP

Table IV. MODELS FOR INFORMATION RETRIEVAL

Implicit feedback models	Concept	Description	Limitation	Evaluation
Binary Voting [4]	**Heuristic based on length of doc representations that assumes useful terms will appear in many of the representations that a searcher chooses to view**	**Each document representation votes for the terms it contains, terms with the highest overall vote were those taken to best describe the information viewed by the searcher and were used to approximate searcher interests**	**Top ranking sentences and title which may be indicative of document content are short and will have low indicativity**	**Best precision wrt search effectiveness**
Jeffery's conditioning [4]	**Jeffery's rule of conditioning assigns a score to each nonstop word term based on normalized frequency of its occurrence**	**Captures uncertain nature of implicit evidence using a measure of confidence to estimate worth of relevant information. Further searchers travels relevance paths not documents**	**Use of representation length may not be always appropriate as a measure of indicativity**	**Best precision wrt search effectiveness, learned at faster rate, highest level of consistency across search topics**
Random term selection [4]	**Assigns a random score b/w 0 and 1 to terms from viewed representations**	**At the end of each relevant path, model ranks the terms based on random scores and uses top scoring terms to expand original query**	**No learning experience**	**Performed poorly, though increased precision over baseline**
Boolean Model [8]	**query is in form of Boolean expression of terms in which terms are combined with operators AND, OR, and NOT**	**Views each document as a set of words**	**It is not possible to obtain a ranking of retrieval results**	**Professionals prefer Boolean query model**
Vector Space / tf-idf Model [6], [8], [13], [14]	**Assigning a score based on query-document pair as vectors Implicitly assumes that terms are pairwise independent Oldest**	**Index and retrieve document by metadata, weigh importance of term based on statistics of occurrence of term**	**If users are unfamiliar to express their information needs with queries, tf-idf cannot provide suitable documents**	**Does not perform well for large document collections in early TREC evaluations**
Topical Relevance Model [6]	**Classification task Types are LSA, PLSA, LDA**	**output from a classifier is the posterior probability of the class membership values since each word can belong to multiple topics with varying membership values**	**LDA is a parametric method, no. of topics have to be preconfigured before inferring posterior probabilities of model**	**Explore variable no. of sentences for improving retrieval effectiveness Hierarchical topic modeling can be applied**
Probabilistic model	**Based on probabilistic distributions of terms in**	**Assumes all documents presented to the model are in**	**The relevance paths were not of sufficient**	**WPQ document model is best for**

- WPQ document , path, ostensive based [4]	**relevant and non relevant documents produces effective terms for query expansion**	**some way relevant. Terms from each complete relevance path are pooled together and ranked. Considers each representation in relevance path separately and ranks each term**	**size and did not contain a sufficient mixture of terms from which WPQ could chose candidates for query expansion. WPQ models need more training to reach a level where terms they recommend appear to match those in relevant distribution.**	**relevant and worst for non relevant retrieval**
Probabilistic model - Binary Independence Model [6]	**Assumes that terms are pairwise independent**	**Information needs converted to query representations and documents to document representations. Probability theory provides reasoning under uncertainty**	**Relies on the boolean presence of a term in a document, and does not use the term frequencies or document length information**	**BM25 outperforms BIM**
Language Model [3,6,8]	**Special case of probabilistic retrieval**	**estimates the posterior probability of generating a document from the query using the complementary prior probabilities of generating a query from a document**	**When it works with microblog retrieval tasks, the document length is short and majority features don't occur more than once thus don't have enough statical information**	**Since tweets are in multiple languages, the model can be applied to applications in different languages**
Language and Relevance model - Query likelihood model [3,8]	**a function that puts a probability measure over strings drawn from some vocabulary is equivalent to probabilistic Finite Automaton**	**Documents are ranked on posterior probability using Baye's rule Query language model may generate the document**	**There is much less text available for estimation based on the query text, and so the model will be worse estimated and will have to depend more on being smoothed with some other language model**	**TBPRF shows av. Improvement over QL**
Relevance Model - Okapi BM25 [3,6]	**BM25 is a probabilistic state-of-the-art retrieval model that can compute the similarity between document and query containing words Extends BIM**	**BM25 model scores a document by accumulating the idf values of the query terms multiplied by the factor of frequency of each term and the document length**	**Doesnot use both posterior and prior probability**	**TBPRF shows av. improvement over BM25**
Relevance Model 3 [1,3]	**Robust model that estimates relevance feedback using QL, BM25 Most effective and robust among query expansion models**	**For the relevance model settings, we set k = 50 pseudo-relevant documents and selected n = 20 feedback terms. Uses irrelevant information when relevant documents are sparse and irrelevance feedback when irrelevant documents are sparse**	**Cannot automatically obtain irrelevance feedback data**	**Most effective and robust among query expansion models**

Table V. INFORMATION RETRIEVAL OPTIMIZATION

Optimization Technique	Description	Future Scope
Genetic Algorithm [9]	**New fitness function for approximate information retrieval**	**Improvements can be made**

	which is very fast and very flexible than cosine similarity fitness function	**in selecting population, fitness function, crossover and mutation points**
Swarm Intelligence (Particle Swarm and Artificial Bee Colony Optimization) [10,11]	**Novel PSO algorithm considering search space as whole collection of documents and inverted file achieved superiority in terms of scalability while yielding comparable quality**	**To work towards increase in response time**
Swarm Intelligence (Ant Colony Optimization) [12]	**Reduce information overload by exploiting collective users' behavior using NaiveRank, RandomRank, SessionRank depending on user's intent**	**To implement in a real time scenario / in an online environment**

V. CONCLUSION AND FUTURE WORK

The author has created a platform to work in the field of Information Retrieval by discussing the various challenges faced by other researchers and various research directions in the field. The author has further compared work w.r.t. Relevance Feedback models – Implicit, Explicit and Pseudo and concludes that out of the three relevance feedback models, Distribution separation of relevant and non relevant documents in Explicit model and query expansion could yield effectiveness in Information Retrieval. The author next analyzes various implicit feedback models in detail which has made her decide that Binary Voting and Jeffery's conditioning provide best precision, professionals prefer Boolean model and WPQ model is best for relevant retrieval. Thirdly the author realized that application of various optimization techniques could bring effectiveness in Information Retrieval, hence compared various Optimization Techniques used by researchers for IR. Lastly author is helping researchers in choosing the datasets for their implementations and various evaluation metrics too. This has led to the conclusion that P@N, NDCG and MAP are mostly used by researchers as evaluation metric and is being experimented mostly on TREC datasets for IR. The author has also tried to find dispersed application areas of research in IR as Patent search, microblog retrieval and scientific data retrieval.

Future work can lie in the field of query optimization through expansion and segmentation, optimized data structures for IR, study of Learning to Rank models, Semantics for IR, personalized or customized retrieval, privacy preserving retrieval, ranking algorithms, web mining, applying Machine Learning to IR, providing retrieved information effectively as feeds, applying mind mapping to IR, query time optimization, etc. Thus, I conclude that this research work and the work in future would definitely help other researchers in carrying out their work.

REFERENCES

[1] Z. Peng, Y. Qian, H. Yuexian, S. Dawei, L. Jingfei and H. Bin, "Distribution separation method using irrelevance feedback data for Information Retrieval", ACM transaction on Information Systems and Technology, http://dx.doi.org/doi:10.1145/2994608, Vol. 9, No. 4, Article 39, Mar 2015.

[2] Zhang P. "Approximating true relevance model in relevance feedback", PhD thesis, 2013.

[3] K. Albishre, Y. Li, Y. Xu, "Effective pseudo relevance for micro blog retrieval", ACSW, Feb 2017, Geelong, Australia. DOI: http://dx.doi.org/10.1145/3014812.3014865.

[4] R. W. White, I. Ruthven, J. M. Jose, C. J. V. Rijsbergen, "Evaluating Implicit Feedback Models Using Searcher Simulations", ACM transactions on Information System, vol 23, no.3, pp 325-361, New York, USA. Jul 2005.

[5] S. Takeuchi, K. Sugiura, Y. Akahoshi, K. Zettsu, "Spatio temporal pseudo relevance feedback for scientific data retrieval", IEEJ transactions on electrical and electronic engineering 2017: 124-131, wiley pub. DOI:10.1002/tee.22352.

[6] D. Ganguly, "Topical Relevance Models", PhD Dissertation, Aug 2013.

[7] F. Shoeleh, M. Azimzadeh, A. Mirzaei, M. Farhoodi, "Similarity based Automatic Web Search Engine Evaluation", IEEE 8th International Symposium on telecommunications, Mar 2017, Tehran, Iran.

[8] C. D. Manning, P. Raghavan, H. Schutze, "Introduction to Information Retrieval": Book, Cambridge University press, 2008, ISBN: 0521865719.

[9] A. A. A. Radwan, B. A. A. Latef, A. M. A. Ali, and O. A. Sadek "Using Genetic Algorithm to Improve Information Retrieval Systems": Book Chapter, Proceedings of World Academy of Science Engineering and Technology: Book, Cairo, Egypt, 2006. ISSN: 1307-6884.

[10] H. Drais, "Web Information Retrieval using Particle Swarm Optimization based approaches", IEEE/WIC/ACM conference on Web Intelligence and Intelligent agent Technology, Oct 2011, Lyon, France. DOI 10.1109/WI-IAT.2011.225.

[11] M. J. Hadi, "Improvement of web Information Retrieval using Swarm Intelligence" Thesis '2013.

[12] A. Malizia, K. A. Olsen, T. Turchi, P. Crescenzi, "An ant-colony based approach for real-time implicit collaborative information seeking", Information Processing and Management 53 (2017) 608-623, ScienceDirect, Elsevier.

[13] K. K. Agbele, E. F. Ayetiran, K. D. Aruleba ; D. O. Ekong "Algorithm for Information Retrieval Optimization", IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference, Nov, 2016, Vancuover, BC, Canada. DOI: 10.1109/IEMCON.2016.7746242.

[14] K. K. Agbele, E. F. Ayetiran, K. D. Aruleba, "Algorithm for Information Retrieval Optimization", International Journal of Computer, Electrical, Automation, Control and Information Engineering, International Science Index, Computer and Information Engineering, International scholarly and scientific research and innovation10(9), World Academy of Science, Engg. & Tech, Vol 10, No. 9, 2016.