Advances in Networks
2017; 5(2): 40-46
http://www.sciencepublishinggroup.com/j/net
doi: 10.11648/j.net.20170502.12
ISSN: 2326-9766 (Print); ISSN: 2326-9782 (Online)

A Survey of Information Retrieval Techniques

Mang'are Fridah Nyamisa, Waweru Mwangi, Wilson Cheruiyot

Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
waweru_mwangi@icsit.jkuat.ac.ke (W. Mwangi), wilchery68@yahoo.com (W. Cheruiyot)

To cite this article:
Mang'are Fridah Nyamisa, Waweru Mwangi, Wilson Cheruiyot. A Survey of Information Retrieval Techniques. *Advances in Networks*. Vol. 5, No. 2, 2017, pp. 40-46. doi: 10.11648/j.net.20170502.12

Received: June 25, 2017; **Accepted**: July 10, 2017; **Published**: November 28, 2017

Abstract: The explosive growth of resources stored in various forms and transmitted over the internet has necessitated researches into information retrieval technologies. The major information retrieval mechanisms commonly employed include vector space model, Boolean model, Fuzzy Set model, and probabilistic retrieval model. These models are used to find similarities between the query and the documents to retrieve documents that reflect the query. These approaches are based on key-word, which uses lists of keywords to describe the information content. In this paper, a survey of these models is provided in order to understand their working mechanisms and shortcomings. This understanding is vital as it facilitates the choice of an information retrieval technique, based on the underlying requirements. The results of this survey revealed that the current information retrieval models fall short of the expectations in one way or the other. As such, they are not ideal for high precision information retrieval applications.

Keywords: Information Retrieval, Model, Fuzzy, Boolean, Probabilistic, Query

1. Introduction

The information retrieval strategies transform documents into suitable representations so that relevant documents can be retrieved effectively. Each of these strategies integrates specific models for the document representation process [1]. These models are further grouped according to dimensions, the mathematical basis (first dimension) and the properties of the model (second dimension).

[2] explain the fact that an information retrieval model governs how a document and a query are represented and how the relevance of a document to a user query is defined. There are four main IR models: Boolean model, vector space model, language model and probabilistic model. The most commonly used models in IR systems and on the Web are the first three models.

Although these three models represent documents and queries differently, they use the same framework. They all treat each document or query as a *bag of words*or *terms*. Term sequence and position in a sentence or a document are ignored. That is, a document is described by a set of distinctive terms. A term is simply a word whose semantics helps remember the document's main themes.

2. Information Retrieval Models

In this section, the various information retrieval models are grouped into two main categories; the first dimension models (mathematical based) and model properties (second dimension). The information retrieval techniques that fall into each of these broad categories are then discussed.

2.1. First Dimension Models

In their study, [3] explain that the first dimension models can further be classified into four categories namely: set-theoretic, algebraic models, probabilistic models and feature-based retrieval models. The set-theoretic models include standard Boolean model, extended Boolean model and fuzzy retrieval. The algebraic models include vector space model, generalized vector space model, (enhanced) topic-based vector space model, extended Boolean model, and latent semantic indexing. The probabilistic models include binary independence model, probabilistic relevance model, uncertain inference, language models, divergence-from-randomness model, and latent Dirichlet allocation. On the other hand, feature-based retrieval models treat documents as vectors of values of attributes. They then search for the best way to combine these attributes into a distinct relevance score.

2.1.1. Standard Boolean Model

The standard Boolean model uses the notion of exact matching to equate documents to the user query. Both the query and the retrieval are based on Boolean algebra. In the Boolean model, documents and queries are represented as sets of terms. That is, each term is only considered present or absent in a document. According to [4], this model is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms.

In their research paper, [5] developed a model that was an improvement of Boolean Information Retrieval (BIR), based Semantic Web (SW) techniques. Their model included an ontology that was merged with the traditional Information retrieval model. It employed ontology-based approach to extract Reference Concept (RC) for each term in the collection and in the query. In so doing, their semantic Boolean information retrieval model achieved high precision in comparison with traditional model.

2.1.2. Extended Boolean Model

In their study, [6] discuss that the Boolean model does not consider term weights in queries moreover, the result set of a Boolean query is often either too small or too big. These shortcomings led to the development of an extended Boolean model.

The aim of this new model was to employ partial matching and term weights as in the vector space model [7]. Conceptually, it combines the characteristics of the Vector Space Model with the properties of Boolean algebra. Additionally, the model ranks the similarity between queries and documents. In so doing, a document may be fairly relevant if it matches some of the queried terms and will be returned as a result, whereas in the Standard Boolean model this was not the case.

However, as [2] explain, this extended model has some setbacks in that the exact matching may retrieve too few or too many documents; it is hard to translate a query into a Boolean expression; all terms are equally weighted; and the fact that this approach is more like data retrieval than information retrieval technique. Moreover, this model cannot rank documents in decreasing order of relevance.

2.1.3. Fuzzy Retrieval Model

The concept of fuzzy logic model allows the manipulation of weights given to terms, and to perform an accumulation on them [8]. This logic permits intermediary truth values to be defined between conventional evaluations of true and false. The fuzzy retrieval is based o fuzzy set, which are sets whose elements contain degrees of association. In this approach, an element either belongs or does not belong to a given set.

According to [9], the fuzzy set theory allows the steady evaluation of the membership of elements in a set. This is facilitated by the help of a membership function valued in the real interval. In this approach, the transition from membership to non-membership could be gradual, instead of it being sudden as is the case of Boolean theory.

The main challenge of the fuzzy retrieval, according to [10], is that there has not been a recognized basis for determining the rating of membership. It is a subjective measure that hinges around the context. Moreover, the set membership does not mean the same thing at the operational level in each and every context. Effectively, degree of membership can have one of the following semantics: a degree of similarity, a degree of preference, or a degree of uncertainty.

2.1.4. The Vector Space Model

Vector Space Model (VSM) is an algebraic model representing textual information as a vector [11]. In this model, after the required initial pre-processing, a dictionary of terms (vocabulary) is extracted from each source document which is then compared against all the suspicious documents. In this approach, the documents and queries are represented as vectors in multidimensional space the dimensions of this vector space are the terms used to build an index to represent the documents.

This model is employed in information retrieval, indexing and relevancy rankings and can be successfully used in evaluation of web search engines. According to [12], the procedure for this model can be divided in to three stages. The first stage is the document indexing where content bearing terms are extracted from the document text. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure.

Ideally, this model represents the importance of a word using term frequency Inverse document frequency (tf-idf) metric. Inverse document frequency, idf(t) is then calculated which emphasizes that a term which is almost present in the entire corpus of documents is not good [12]. Finally their product, i.e. tf-idf is calculated and the similarity between document vectors is calculated using cosine similarity.

[13] explain that the VSM method gives good results when used along with approaches like ranking of documents, Latent Semantic Indexing. This model is perhaps the best known and most widely used IR model. A document in the vector space model is represented as a weight vector, in which each component weight is computed based on some variation of TF or TF-IDF scheme. As [14] elaborate, the VSMs have several attractive properties: VSMs extract knowledge automatically from a given corpus, thus they require much less labour than other approaches to semantics, such as hand-coded knowledge bases and ontologies. Additionally, it is a simple model based on linear algebra; it employs term weights instead of binary; it permits computing a continuous degree of similarity between queries and documents; it permits ranking of documents according to their possible relevance; and can consent to partial matching.

The main setbacks of the vector space model, as noted by

[15] are that : Long documents are poorly represented since they have poor similarity values that are in form of a small scalar product and a large dimensionality; the search keywords must precisely match document terms, that is , word substrings might result in a false positive match; it is prone to semantic sensitivity in that documents with similar context but different term vocabulary would not be associated, resulting in a false negative match; the order in which the terms appear in the document is lost in the vector space representation; it theoretically assumes that the terms are statistically independent; and the weighting is intuitive but not very formal.

2.1.5. Generalized Vector Space Model

In their study, [16] noted that one of the most accepted retrieval models to establish similarity between documents and queries is the Vector Space Model (VSM). Typically this model assumes pairwise orthogonality among the vectors representing the index terms. The consequence of this is that the index terms are independent of each other. Therefore, the vector space model does not take into consideration the idea that two index terms can be semantically related. For this reason, the Generalized Vector Space Model (GVSM) has been introduced. In this model, the index terms are composed of smaller elements and term vectors are not considered pair-wise orthogonal in general.

According to [17] before utilizing the semantic relatedness, the document contents must be annotated via the Named Entity Linking (NEL). Basically, NEL involves the identification of mentions of named entities in a text and linking them to the corresponding entities in a knowledge base. A number of NEL approaches exist and most of them incorporate natural language processing, such as named entity recognition, co-reference resolution, and word sense disambiguation (WSD) with statistical, graph-based, and machine learning techniques.

However, [18] explain that none of these approaches provides an all-round service for end-user centered semantic search, which concurrently builds on a theoretically sound retrieval model and is proven to be practically useful. Moreover, all these approaches do not take advantage of the relationships of concepts represented in a document.

2.1.6. Enhanced Topic-Based Vector Space Model

This model has the capability of representing linguistic phenomena using a semantic ontology. In their study, [19] explain that spam has grown to be a serious issue in computer security. This is because it acts as a channel for threats such as computer viruses, worms and phishing. Empirically, they explain that more than 85% of received e-mails are spam messages.

The traditional approaches to prevent these messages involve simple techniques such as sender blacklisting or the use of e- mail signatures. However, as [20] notes these techniques are no longer completely reliable. As such, many current solutions utilize machine-learning algorithms trained using statistical representations of the terms that usually appear in the e-mails. However, these methods are merely syntactic and are incapable in accounting for the underlying semantics of terms within the messages.

In their research, [21] explored the use of semantics in spam filtering by representing e-mails with a recently introduced Information Retrieval model: the enhanced Topic-based Vector Space Model (eTVSM). Based upon this representation, several well-known machine-learning models were applied. The results demonstrated that the proposed method could detect the internal semantics of spam messages.

2.1.7. Latent Semantic Indexing

In his research, [22] discuss that basically information is retrieved by exactly matching terms in documents with those of a query. However, these lexical matching methods are imprecise when employed to match a user's query. This emanates from the fact that there are generally many ways to express a given concept, by use of synonyms. In some circumstances, the literal terms in a user's query may not match those of a relevant document. Additionally, many phrases and words have multiple meanings, poly-semy. This means that terms in a user's query will literally match terms in irrelevant documents.

According to [23] an enhanced approach would allow users to retrieve information on the basis of a conceptual topic or meaning of a document. This is where latent semantic indexing comes handy. This is a technique in natural language processing of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

This approach has been proven very effective in categorization of content into predefined concepts or topics [24]. Ideally, the concepts contained in the documents being categorized are compared to the concepts contained in the example items, and a category (or categories) is assigned to the documents based on the similarities between the concepts they contain and the concepts that are contained in the example documents.

2.1.8. Binary Independence Model

This is a probabilistic information retrieval technique that treats documents as binary vectors. This means that only the presence or absence of terms in documents is documented. The basic assumption is that the terms are independently distributed in the set of relevant documents and they are also independently distributed in the set of irrelevant documents. According to [25], this representation is an ordered set of Boolean variables, where representation of a document or query is a vector with one Boolean element for each term under consideration.

In this approach, a document is represented by a vector:

$$\text{doc} = (x_1, ..., x_m) \quad (1)$$

where $x_t=1$ if term t is present in the document *doc* and $x_t=0$ if it's not.

In their study, [26] explain that queries are represented in a similar way, where autonomy indicates that the terms in the document are considered independently from each other and no association between terms is therefore modeled.

According to [25], this postulate is very restrictive in as much as it gives better results for many situations. This independence is the raw assumption of a Naive Bayes classifier, where properties that imply each other are even so treated as independent for the sake of simplicity. Arguably, this assumption permits the representation to be treated as an instance of a Vector space model. This is done by considering each term as a value of 0 or 1 along a dimension orthogonal to the dimensions used for the other terms.

2.1.9. The Probabilistic Relevance Model

The ultimate objective of a retrieval model is to measure the degree of relevance of a document with respect to the given query. Probabilistic models are widely used to measure the likelihood of relevance of a document by combining within document term frequency and term specificity in a formal way [27]. Recent research shows that term frequency (*tf*) normalization that factors in multiple aspects of term salience is an effective scheme. However, existing models do not fully utilize these *tf* normalization components in a principled way. Moreover, most state of the art models ignore the distribution of a term in the part of the collection that contains the term.

In their paper, [28] introduced a new probabilistic model of ranking that addresses the above issues. They argued that, since the relevance of a document increases with the frequency of the query term, this assumption can be used to measure the likelihood that the normalized frequency of a term in a particular document will be maximum with respect to its distribution in the elite set. Thus, the weight of a term in a document is proportional to the probability that the normalized frequency of that term is maximum under the hypothesis that the frequencies are generated randomly. To that end, they introduced a ranking function based on maximum value distribution that uses two aspects of *tf* normalization. The merit of the proposed model was demonstrated on a number of recent large web collections. The results obtained showed that the proposed model outperformed the state of the art models by significantly large margin.

Moreover, in their research paper, [29] pointed out that the probabilistic model determines the relevance between queries and documents it retrieves. This model is intended to estimate the probability that the required documents are relevant to the query term. The procedure adopted in this model is as follows: The user issues a short and simple query; the search engine returns a set of documents; the user marks some documents as relevant, while he marks some of these documents as irrelevant; the search engine computes a new representation of the information need. Ideally, this new computation should be better than the initial query. The search engine runs new query and returns new results. The new results have hopefully better recall compared to the first one.

However, as [30] pointed out, various setbacks are inherent in the probabilistic model: it does not categorize documents based on their relevance; does not consider document weighing in building the index term and freedom of supposition for the index term; generally, the idea of probabilistic model is within a probabilistic scope, which allows the user to retrieve and which documents are relevant and which documents are not. This process needs to be repeated until a certain stage where a set of answer patterns can be used to describe the query made. Unfortunately, the probability that those retrieved documents are relevant to the query cannot be computed, making it one of the weaknesses in the probabilistic model.

A further weakness identified in the probabilistic model, according to [31] is how this model has set the relevancy of documents by only assuming the weights of relevancy with binary data, which are not compliant to the frequency of the index terms, which appear in each document.

2.1.10. Uncertain Inference Model

This model deals with ways of formally defining query and document relationship in Information retrieval, as a mechanism of deriving consequences from human knowledge through uncertain set theory [32]. In this approach, the query supplied by the user is treated as a set of assertions about the desired document. The information retrieval system's main activity is then to deduce, given a particular document, if the query statements are true. In situations where the statements are true, the document is retrieved.

[33] explain that the main challenge of this approach is that the contents of documents are not sufficient to assert the queries. Therefore, a knowledge base of facts and rules is required. However, some of them may be uncertain since there may be a probability associated to using them for inference.

Other models include the language model, Divergence-From-Randomness Model, and Latent Dirichlet Allocation model. Language modeling is an elementary task in artificial intelligence and natural language processing (NLP). In their study, [31] elaborate that a language model finds crucial applications in speech recognition, text generation, and machine translation. A language model is formalized as a probability distribution over a sequence of strings or words, and traditional methods usually involve making an *n-th* order Markov assumption and estimating *n-gram* probabilities through counting and subsequent smoothing.

According to [34] one merit of count-based models is that they are simple to train. On the other hand, probabilities of rare *n-grams* can be poorly estimated due to data sparsity, despite smoothing techniques the intention of language modeling is to estimate the probability distribution of various linguistic units, such as words or sentences. Count-based n-gram language models were among the first techniques in this field. This approach's intention is to assign the probability distribution of a given word observed after a fixed number of previous words.

In their study, [35] point out that feed-forward neural language model has been developed that can achieve substantial improvements in perplexity over count-based language models. Further, [36] showed that this neural language model could simultaneously learn the conditional probability of the latest word in a sequence as well as a vector representation for each word in a predefined vocabulary.

[37] pointed out that when modeling a corpus, these language models assume the mutual independence among sentences, and the task is often reduced to assigning a probability to a single sentence. The challenges of language models are that typically, the *n*-gram model probabilities are not derived directly from the frequency counts. Therefore, models derived this way have severe problems when confronted with any *n*-grams that have not explicitly been seen before. Another setback is that of global information access. This means the requirement to satisfy human information needs through natural, efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language.

Additionally, [38] illustrated that there is the issue of contextual retrieval which requires that combine search technologies be combined with knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs. Lastly, this approach requires the development of intelligent classification algorithms that will be able to unobtrusively elicit user feedback, combine it with contextual and historical evidence, and produce effective structured annotation of new data.

According to [33], Divergence-From-Randomness Model is one type of probabilistic model where term weights are calculated by determining the divergence between a term distribution produced by a random process and the actual term distribution. This information retrieval model employs term-document matching function that are computed by taking the product of two divergence functions.

Latent Dirichlet Allocation is a generative statistical model that permits sets of observations to be explained by groups not observed so as to provide an explanation as to why some parts of the data are analogous [39]. In this approach, each document is treated as a mixture of various topics. It also assumes that there are k underlying latent topics according to which documents are generated, and that each topic is represented as a multinomial distribution over the |V| words in the vocabulary. Therefore, a document is generated by sampling a mixture of these topics and then amalgamating words from that mixture.

2.2. Second Dimension Models

These models include models without term-interdependencies, models with immanent term interdependencies and models with transcendent term interdependencies. The first model view various terms as being autonomous while the second model permits for the representation of interdependencies between phrases and terms [40]. However, the degree of the interdependency between two terms is described by the model itself. The last model on the other hand permits depiction of interdependencies between phrases and terms but do not explicitly state how these phrases and terms are related.

3. Challenges with Information Retrieval Models

Each of the information retrieval techniques possess shortcomings in the way it handles the user's query and return query results. For instance, the Boolean model does not consider term weights in queries. In addition, the result set of a Boolean query is often either too small or too big. These shortcomings led to the development of an extended Boolean model. The aim of this new model was to employ partial matching and term weights as in the vector space model. Conceptually, it combines the characteristics of the Vector Space Model with the properties of Boolean algebra.

Additionally, the model ranks the similarity between queries and documents. In so doing, a document may be fairly relevant if it matches some of the queried terms and will be returned as a result, whereas in the Standard Boolean model this was not the case. However, this extended model has some setbacks in that the exact matching may retrieve too few or too many documents; it is hard to translate a query into a Boolean expression; all terms are equally weighted; and the fact that this approach is more like data retrieval than information retrieval technique. Moreover, this model cannot rank documents in decreasing order of relevance.

The main setbacks of the vector space model are that : Long documents are poorly represented since they have poor similarity values that are in form of a small scalar product and a large dimensionality; the search keywords must precisely match document terms, that is, word substrings might result in a false positive match; it is prone to semantic sensitivity in that documents with similar context but different term vocabulary would not be associated, resulting in a false negative match; the order in which the terms appear in the document is lost in the vector space representation; it theoretically assumes that the terms are statistically independent; and the weighting is intuitive but not very formal.

On its part, the probabilistic model does not categorize documents based on their relevance; does not consider document weighing in building the index term and freedom of supposition for the index term; generally, the idea of probabilistic model is within a probabilistic scope, which allows the user to retrieve and which documents are relevant and which documents are not. This process needs to be repeated until a certain stage where a set of answer patterns can be used to describe the query made. Unfortunately, the probability that those retrieved documents are relevant to the query cannot be computed, making it one of the weaknesses in the probabilistic model. A further weakness identified in the probabilistic model is how this model has set the relevancy of documents by only assuming the weights of relevancy with binary data, which are not compliant to the frequency of the index terms, which appear in each document.

The challenges of language models are that typically, the n-gram model probabilities are not derived directly from the frequency counts. Therefore, models derived this way have severe problems when confronted with any n-grams that have not explicitly been seen before. Another setback is that of global information access. This means the requirement to satisfy human information needs through natural, efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language.

Additionally there is the issue of contextual retrieval which requires that combine search technologies be combined with knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs. Lastly, this approach requires the development of intelligent classification algorithms that will be able to unobtrusively elicit user feedback, combine it with contextual and historical evidence, and produce effective structured annotation of new data. These are the challenges that this research sought to address.

4. Conclusions and Recommendations

Various shortcomings have been observed in the current information retrieval models. As such, novel information model based on n-grams approach is suggested. The model should have the capability of semantically indexing data according to concepts by utilizing n-grams to retrieve these documents. In this approach, documents whose unigram language models are similar to the query's unigram language model are more likely to be relevant. The performance of the model needs to be evaluated using statistical metrics such as precision, recall and F-Measure. The search technique employing n-grams is assured to retrieve relevant documents since all the semantically related information to the user query would be taken into consideration in the new design. This new search techniques operational criteria are crucial to database designers in that it would help them design their databases in such a way that all the relevant information regarding the users' query are returned to the user. This effectively overcomes the keyword-based approach where the search results are bound to the keyword supplied by the user.

References

[1] B. Jansen and S. Rieh (2010). *The Seventeen Theoretical Constructs of Information Searching and Information Retrieval*. Journal of the American Society for Information Sciences and Technology. 61(8), 1517-1534.

[2] I. Sutskever, O. Vinyals and Q. Le (2014). *Sequence to Sequence Learning with Neural Networks.*

[3] M. Sanderson and W. Bruce (2012). The History of Information Retrieval Research. Proceedings of the IEEE. 100: 1444–1451.

[4] R. Baeza, and B. Ribeiro (2011). *Modern Information Retrieval*: Second edition. Addison-Wesley, New York, NY, USA.

[5] E. Elabd, E. Alshari, and H. Abdulkader (2014). *Semantic Boolean Arabic Information Retrieval*. The International Arab Journal of Information Technology.

[6] Q. Shatnawi B. Yassein B. and R. Mahafza (2012). *A Framework for Retrieving Arabic Documents Based on Queries Written in Arabic Slang Language*. Journal of Information Science, vol. 38, pp. 350-365.

[7] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil (2014). *A Latent Semantic Model with Convolutional-pooling Structure for Information Retrieval*. In Proceedings of CIKM.

[8] R. Harastani (2010). *Information Retrieval With Fuzzy Logic*. Texmex.

[9] W. Onifade and J. Ibitoye (2016). F*uzzy Latent Semantic Query Expansion Model for Enhancing Information Retrieval.* University of Ibadan, Nigeria.

[10] B. Yates and R. Neto (2012). *Modern information retrieval*. Addison Wesley, 2011.

[11] D. Turney, and P. Pantel (2010). *From Frequency to Meaning: Vector Space Models of Semantics*. Journal of Artificial Intelligence Research.

[12] N. Singh andK. Dwivedi (2012*). Analysis of Vector Space Model in Information Retrieval*. National Conference on Communication Technologies & its impact on Next Generation Computing.

[13] R. Kiros, Y. Zhu, R. Salakhutdinov, S. Zemel, A. Torralba, R. Urtasun, and S. Fidler (2015). *Skip-thought vectors.*

[14] R. Pascanu, C. Culcehre, K. Cho, and Y. Bengio, (2013). *How to Construct Deep Neural Networks*.

[15] M. Dragoni, Celia da Costa Pereira, G. B Andrea. Tettamanzi, (2012). *A Conceptual Representation of Documents and Queries for Information Retrieval System using Light Ontologies*. Expert Systems with Applications pp. 10376–10388, Elsevier.

[16] C. Exeler and H. Sack (2015). *Linked Data Enabled Generalized Vector Space Model To Improve Document Retrieval.* Hasso-Plattner-Institute for IT-Systems Engineering.

[17] R. Usbeck (2015). *GERBIL: general entity annotation benchmark framework.* In 24th WWW conference.

[18] T. Tietz, J. Waitelonis, J. Jager, and H. Sack (2014). *Smart media navigator: Visualizing recommendations based on linked data*. In 13th International Semantic Web Conference, Industry Track, pages 48{51}.

[19] I. Santos, B. Sanz C. Laorden and G. Bringas (2012). Enhanced Topic-based Vector Space Model for semantics-aware spam filtering. Expert Systems with Applications 39:437-444.

[20] H. Drucker (2013). Support Vector Machines for Spam Categorization.

[21] M. Kwak and G. Leroy (2013). Development and Evaluation of a Biomedical Search Engine using a Predicate-based Vector Space Model.

[22] S. Clark (2013). *Topic Modelling and Latent Dirichlet Allocation.* Machine Learning for Language Processing.

[23] D. Blei (2012). *Probabilistic topic models*. Communications of the ACM, 55(4):7784.

[24] S. Liangcai B. Long, M. Weiyi (2014). A Latent Topic Model for Complete Entity Resolution. 25th IEEE International Conference on Data Engineering.

[25] B. Stefan L. Charles V. Gordon (2014). Information Retrieval: Implementing and Evaluating Search Engines. MIT Press.

[26] D. Manning P. Raghavan S. Hinrich (2013). Introduction to Information Retrieval. Cambridge University Press.

[27] H. Paik, (2013). *A novel TF-IDF weighting scheme for effective ranking*. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland.

[28] R. Cummins, H. Jiaul, L. Yuanhua, A. Pólya (2015). *Urn Document Language Model for Improved Information Retrieval*. ACM Transactions on Information Systems (TOIS), v.33 n.4, p.1-34.

[29] P. Sojka and H. Schütze (2015). *Introduction to Information Retrieval*. Faculty of Informatics, Masaryk University.

[30] Y. Baeza, R. Ribeiro (2011). *Modern Information Retrieval.*

[31] Y. Kim, Y. Jernite, D. Sontag, M. Rush (2016). *Character-Aware Neural Language Models*. School of Engineering and Applied Sciences Harvard University.

[32] P. Wise, M. Henrion (2013). A Framework for Comparing Uncertain Inference Systems to Probability. Cornell University Library.

[33] E. Kyburgand, C. Teng (2015). Uncertain Inference.

[34] S. Zhang, H. Jiang, M. Xu, J. Hou, and L. Dai (2015). *The Fixed- Size Ordinally-Forgetting Encoding Method for Neural Network Language Models*. In Proceedings of ACL.

[35] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocky (2011). *Empirical Evaluation and Combination of Advanced Language Modeling Techniques.* In Proceedings of INTERSPEECH.

[36] M. Sundermeyer, H. Ney, and R. Schluter (2015). *From feedforward to recurrent lstm neural networks for language modeling*. Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23(3):517–529.

[37] S. Goldwater (2015). *Introduction to Computational Linguistics: N-gram language models.*

[38] D. Matthew(2012). *Adadelta: An adaptive learning rate method*.

[39] G. Amati (2015). Divergence from Randomness Models.

[40] S. Hinrich (2011). Introduction to Information Retrieval. Institute for Natural Language Processing, Universit¨at Stuttgart.