Journal of Theoretical and Applied Information Technology
15th June 2019. Vol.97. No 11
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LANGUAGE MODEL FOR DIGITAL RECOURSE OBJECTS RETRIEVAL

[1]WAFA' ZAAL ALMAA'ITAH, [2]ABDULLAH HJ TALIB,

[3]MOHD AZAM OSMAN, [4]ADDY ALQURA`AN

[1] School of computer sciences, Universiti Sains Malaysia, Penang, Malaysia
Department of Basic Sciences, Hashemite University, Jordan
[2] School of computer sciences, Universiti Sains Malaysia, Penang, Malaysia
[3] School of computer sciences, Universiti Sains Malaysia, Penang, Malaysia
[4] Department of Basic Sciences, Hashemite University, Jordan
E-mail: [1]wafaa_maitah@hu.edu.jo, [1]wzma15_com083@student.usm.my, [2] azht@usm.my,
[3]azam@usm.my, [4] addy@hu.edu.jo

ABSTRACT

Language model has been successfully applied for use in information retrieval to retrieve structure and unstructured information. Typically, language model involves three basic models namely: N-gram language models, smoothing model and estimation model. Language model has been approved outperforms of other retrieval model such as vector space model and probabilistic model. The problem arises when language model uses to retrieve digital Resource Objects which use metadata to describe their content. Digital Resource Objects have special three characteristics: lack in metadata content (short document), short query, and heterogeneity metadata content. This paper presents a performance comparison among information retrieval models (Vector Space Model and Probabilistic Model) using a Digital Resource Objects (CHiC2013 collection). Further, an overview for language model approaches to determine which models are suitable for digital Resource Objects, despite being a traditional review, a comprehensive comparative analysis is conducted among different approaches of Language model.

Keywords: L*anguage Model, Information Retrieval, Digital Recourses Object*

1. INTRODUCTION

The language model (LM) finds applications in a wide gamut of fields such as speech recognition, natural language processing [1, 2], information retrieval [3, 4] and machine translation [5]. From the perspective of information retrieval, the LM projects the word distribution in an input language. A document is normally seen as an example from a language model that lies beneath. That is to say, the document is just one possible account of the knowledge being expressed by the writer; the words used in the set are produced with particular possibilities. These documents are graded by the possibility that every document language model might have produced the query terms of the user.

Several variations of the LM methodology pertaining to information retrieval have been recommended, such as numerous Bernoulli models [6], relevance models [7] and multinomial models (Zhou & Liu, 2008). An LM for information retrieval encompasses the following constituents: (i) A suite of document language models for every document in the collection as well as a suite of query models, (ii) A probability distribution function that allows estimating the likelihood, (iii) A rank function that integrates these produced possibilities for grading the documents with respect to the query. Many noted research works on the LM methodology by Xu, et al. [8], Lavrenko, et al. [9], Xu and Croft [10], Si, et al. [11] have showed that the LM methodologies are quite an effectual probabilistic framework for retrieving information. Bennett, et al. [12] proved that the LM is far better compared to other IR models like the probabilistic model and vector space model.

The main motivation of this research comes from the need for a more effective IR system that enriches and handles DRO content for non-expert users. Therefore, there is a need for better and effective models that can be incorporated in IR to allow the user to access and explore the information on DRO.

Typically, LM involves three basic models namely: N-gram models, smoothing models and estimation models. Figure 1 shows the taxonomy chart of LM approaches in information retrieval. The rest of this paper is organized as follows: Section 2 presents an overview of the N-gram language models. Smoothing models are discussed in Sections 3. The language model estimation methodologies are discussed in section 4. The analysis and observations are presented in section 5. Section 6 presents the conclusion.

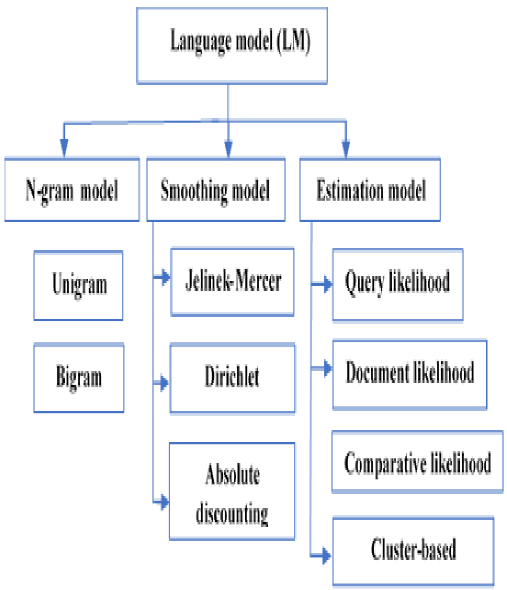

Figure 1. A taxonomy of language model

2. N-GRAM LANGUAGE MODELS

N-gram models are the primary language model form and it is the means by which a possibility is distributed over whole sentences or paragraphs. The fundamental concept is to regard the form of a corpus, text, or language as the possibility of several words appearing in sequence or alone. The unigram models are the general and simplest models whereby the terms are regarded in isolation [13], and so is the bigrams model which considers the terms dependently [14]. Thus, the N-gram models include but are not limited to bigrams and unigrams, N-gram models of higher order are also utilised and include the trigram model [15]. The most straightforward and natural mode of estimating probabilities is the maximum likelihood estimation (MLE) method [16].

2.1 Unigram language model

The unigram model assumes that every word is distributed independent of its history by ignoring all previous words. Also, it's called "bag of words" models, because they assign the same probability to a group of words, the group is a bag that contain a set of words, where each group represent a document with ignoring the order of words. Several approaches have been proposed under the assumption of term independency such as [17-19]. Furthermore, Unigram models are easy to understand, simple models, and they give good results in several information retrieval studies by Westerveld and Vries [20] ,Vulić, et al. [21], Peetz, et al. [22], Yu [23], Choi, et al. [24], Baumel, et al. [25]. In addition, Symonds, et al. [26] has shown that the unigram relevance model is outperformed by dependency N-gram model. Unigram models most common in information retrieval even bigram models have been used.

2.2 Bigram language model

The bigram model supposes that the conditional possibility of the term relies only on the preceding term and is known as a Markov assumption, in which the criteria is that the presence of the following term relies on the preceding term or terms [14]. Markov models are the category of probabilistic models which presume that we can estimate the probability of certain future unit without going too much into the past.

Many variations on the bigram model have been proposed such as Jiang, et al. [27], Nallapati and Allan [28], Eguchi and Croft [29], Bruck and Tilahun [30] and Jiang, et al. [31]. Usually, the likelihood of the appearance of a term relies only on the possibility of the n preceding terms that extend 3-grams, 4-grams, 5-gram, and so on. To make N-gram models useful, it must apply a smoothing to eliminate zero counts [32]. The summary of unigram and bigram methods is shown in Table 1.

Table 1. Summary of unigram and bigram methods

N-gram models	Dependency	Probability estimation	Usage
Unigram	Ignoring the context	order of words not important	Simple counting
Bigram	Depending on the past word	order of words is important	Complicated counting

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. SMOOTHING MODELS

Smoothing is an essential characteristic of language models, as it balances the possibilities of the terms by not counting the term probabilities viewed in the text and regulating zero or low possibilities upwards for other words [33-35]. This evades the problems with zero occurrences, where the words of the query that are not present in a text would cause otherwise a query possibility of zero on the whole. It alters the maximum probability language model predictor to be more efficient. It plays two important roles: 1) enhances the affectivity of the language model. 2) Helps the creation of words that are general and non-informative. To make n-gram models useful must apply a smoothing method that eliminates zero counts. There are many methods available for smoothing, the basic three smoothing methods are: Jelinek-Mercer smoothing, Dirichlet smoothing and Absolute discounting smoothing.

3.1 Jelinek-Mercer Smoothing Model

The Jelinek-Mercer (JM) smoothing put forward by Jelinek and Mercer [36] requires a linear interpolation of the collection model and the maximum probability model, with a coefficient λ. It integrates the query term's relative occurrence in the text D with the term's relative occurrence in the entire collection. In previous works done by Lafferty and Zhai [37] as well as Zhai [38], they have discovered that the quantity of λ 0.1 is appropriate for small queries, and greater values of $\lambda \approx$ 0.7 are more appropriate for lengthy queries. Ding and Wang [39] expanded the JM smoothing method by combining a document-dependent factor to regulate the effect of the collection model and the maximum probability model. Zhai and Lafferty [34] have mentioned that the method of JM smoothing is the poorest for small queries, but the most excellent and more efficient in case of lengthy queries. Smucker and Allan [40] and Losada and Azzopardi [41] showed a significant performance for JM smoothing when documents are long.

3.2 Dirichlet Smoothing Model

The Dirichlet smoothing method makes smoothing reliant on the size of the text. In this model, it is more likely that less smoothing is required. If we employ the multinomial distribution to signify a language model, this distribution's conjugate prior will become the Dirichlet distribution [42]. As μ value becomes smaller, so does the collection model, and the weighting of the relative term is given more emphasis. [43] as well as Zhai and Lafferty [34] have mentioned that the best possible preceding value for μ is about 2,000. In He and Ounis [44], the method of Dirichlet smoothing was remodelled on the basis of the measurement of the correlation among the normalized term occurrence and the length of the document for a query term given. Zhai [45] has shown that the long documents are impacted less by μ and should be tuned or pick average document length. Dirichlet smoothing language model (LMD) is generally considered to be more effective than other smoothing based language models, especially for short queries [46]. Moreover, Azzopardi and Losada [47] as well as Losada and Azzopardi [41] have shown that the method of Dirichlet smoothing is likely to get several small texts and a few amount of long texts.

3.3 Absolute Discounting Smoothing Model

Ney, et al. [48] illustrated a method of smoothing where every non-zero count is reduced by deducting a constant value δ from every term's counts. The possibility mass obtained from the terms that are present is distributed evenly over the unseen events. This technique is similar to Jelinek-Mercer Zhai [38] smoothing, the different being that it reduces the possibility of the seen word by deducting a constant value rather than multiplying it. Lafferty and Zhai [37] as well as Bennett, et al. [12] have mentioned that for small queries, the Dirichlet smoothing method is superior to absolute discounting which in turn is superior to Jelinek-Mercer smoothing technique. For lengthy queries, the JM smoothing method is superior to the Dirichlet smoothing method as well as the absolute discounting method. The summary of smoothing methods is presented in Table 2.

Table 2. Summary of smoothing methods

Smoothing methods	Parameter used	Parameter Dependent	Optimal parameter values	Suitable for
Jelinek-Mercer Smoothing	λ	Interpolate linear and depend on linear weight	0.1for short query 0.7 for long query	Long document Long query
Dirichlet Smoothing	μ	Term frequently and depend on	2000 Constant value	Short document Short

		document		query
Absolute Discounting	δ	Interpolate linear and depend on linear weight	≈ 1	Long document Long query

4. ESTIMATION MODELS

language model for information retrieval can be categorized into four approaches based on retrieval probability estimation method : (i) The approach of query probability retrieval, which is ranked on the basis of the probability of a language model of a document producing a query, (ii) The approach of document possibility, which is ranked on the basis of the possibility of a language model of query producing a document, (iii) Comparative approach, which is ranked on the basis of the possibility of the query being seen as a document's translation, and (iv) Cluster-based language models.

4.1 Query Likelihood Retrieval Model

First proposed by Ponte and Croft [4] and described by Berger and Lafferty [49]. The basic idea behind query likelihood retrieval model is to infer a LM for each document, estimate the probability of query in document, and rank documents based on the probability of a query being generated from a document P(d|q) [50]. To estimate P(d|q) by using the Bayes rule with these assumptions: (i) P(q) is the same for all documents and (ii) P(d) is treated as uniform across all d, (vi) all words are independent. According Zhai and Lafferty [51] the query likelihood model has generalized to the Kullback-Leibler (KL) divergence scoring method, by modelling the query separately. Among many approaches of LM have proposed, the most popular and fundamental one is the query-likelihood language model, it is shown to be theoretically superior and confirmed experimentally by Bruza and Song [52], Mei, et al. [53] , Lv and Zhai [54] and Lin and Bilmes [55]. Furthermore, Cummins, et al. [56] has shown that the query likelihood model with Dirichlet smoothing can be implemented as effectively as traditional retrieval.

4.2 Document Likelihood Retrieval Model

Proposed by Hofmann [57] and modelled by Song and Croft [32] who inversing the direction of the query likelihood approach. It constructs a query language model and computes the probability of the documents being produced using this model. The main process to estimate a document's language model [58]: (i) tokenize and split the document text into terms, (ii) Count the number of times each term occurs, (iii) Count the total number of term occurrences, and (iv) Assign term a probability. According to Lavrenko, et al. [9] the disadvantage of this approach, it has low performance with short query. Due to the queries are often very short [59], the models derived from the short queries are relatively poor. For small and heterogeneous documents, this technique is not considered effective, a fact which has been examined by Spitters and Kraaij [60] whereby they demonstrated that the possibility of producing a document is likely to be smaller for lengthy texts compared to that for small texts and it requires normalization as the texts are of varying lengths.

4.3 Comparative Model

Recommended by [61], Lafferty and Zhai [37] as well as Zhai [43] have created a structure for minimization of risk on the basis of the Bayesian theory of decision. In this model, queries and texts are structured using the LM method; retrieval is considered as a problem for minimization of risk. The resemblance among a query and a text is quantified by the method of Kullback-Leibler (KL) divergence among the query model and the text model [62]. In this framework, the document LM can be predicted like the query probability model; nevertheless, the concern again (similar to that with the document probability model) is to predict for the query a good LM.

4.4 Cluster-Based Language Model

Cluster-based model for language employs the clustering of texts to arrange the collections on the basis of the subjects. Each cluster is supposed to be representing a subject and the model for language can be developed for a particular cluster. Liu and Croft [63] have integrated cluster-based models for language into IR models by substituting text D with the cluster C, P (Q|D) to P (Q|C) to get ranked clusters. Other language model based on clustering includes Zhang, et al. [64] in which the scholar structured cluster production by employing a Dirichlet process mixture framework, in which the base distribution can be considered as the prior of the common English model and the precision factor which regulates the random production procedure for obtaining new clusters. In other work the author Tan, et al. [65] considered each document of a collection is again viewed as a sample and the vocabulary of the corpus as a generated text process. It can compare the distance between two

Journal of Theoretical and Applied Information Technology
15th June 2019. Vol.97. No 11
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

documents perspective. Several studies Caropreso, et al. [66], Tan, et al. [67], Tombros, et al. [68], Bassiou and Kotropoulos [69] have used with bigrams language models and shown that the cluster-based language models could improve the effectiveness of information retrieval. It is evident that the search result clustering improves the experience of the user and the quality of the search results. Dreyfuss, et al. [70], Erkan [71], Mahmoodi and Mansoori [72], Momtazi and Klakow [73] have investigated the effectiveness of cluster-based language model. Although Hearst [74] illustrated the weakness of the clustering in the heterogeneous and hierarchical metadata, and it doesn't yield improvements in IR performance. The advantage of the model, it can obtain ranked clusters. But it has two limitations [75]: (i) it must be used in entirety of document collection and need to deal with a very large corpus, so the process must be fast enough, and (ii) it considers the whole cluster as a big document and it is sometimes impossible for the users to browse the whole documents of relevant clusters. The advantages and disadvantages of estimation models are presented in Table 3.

Table 3: Advantages and disadvantages of estimation models

Estimation Models	Advantage	Disadvantage
Query Likelihood Retrieval Model	• Consistent with short query and long heterogeneous documents. • Straightforward probabilistic retrieval model which integrates the term occurrence directly.	• No Smoothing • Difficulties dealing with related feedback, expansion of query, controlled queries
Document Likelihood Retrieval Model	• consistent with long query • relevance feedback possible	• Inconsistent with short heterogeneous document collections
Comparative Model	• consistent with long short query • relevance feedback possible	• Inconsistent with short heterogeneous document collections
Cluster Based Language Model	• consistent with long and short query • relevance feedback possible	• Inconsistent with heterogeneous document collections

5. EXTRA REFERENCES

For more studies on LM, Table 4 summarizes some of these studies, explaining the N-gram models, estimation models, and smoothing models used in each study, as well as the finding in each study.

Table 4: Language Model Literature

Authors (year)	N-gram model	Estimation model	Finding
Lavrenko and Croft (2001)	Unigram	Relevance Model	An extension to LM by considering the concepts implied by both the query and words in the document.
Cao et al. (2005)	Unigram/ Bi-gram	Term weight	A dependency LM by integrating two types of relationship Extracted from WordNet and co-occurrence relationships.
Lv and Zhai (2009)	Unigram	Term weight	A positional LM that implemented both heuristics in a unified language model.
Kurland and Krikon, (2011)	Unigram	Query likelihood	A LM approach to ranking query-specific clusters by the presumed percentage of relevant documents that they contain.
Benderskyand Croft	Unigram	Query likelihood	A LM retrieval framework that models

(2012)			dependencies between arbitrary query concepts using a query hypergraph.
Yan et al. (2013)	Unigram	Query likelihood	A unified proximity that combined both semantic and positional proximity heuristics to improve the effect of language model smoothing.
Cummins et al. (2015)	Unigram	Query likelihood	A smoothed Polya document language model incorporates word only into the document model.
Momtazi and Klakow (2015)	Unigram	Maximum likelihood	A language models to improve the performance of sentence retrieval in question answering.
Raviv et al. (2016)	Unigram	Maximum likelihood	An entity-based language model which considers both single terms in the text as well as term sequences marked as entities by an existing entity-linking tool.

6. ANALYSIS AND FINDINGS

Besides the LM is used to retrieve unstructured documents and outperforms its counterparts models in IR as mentioned in section 1, it is also suitable for retrieving structured documents in DRO but after adjusting its process [76] due to the special characteristics that DRO have: lack in metadata content (short document), short query, and heterogeneity metadata content.

Despite N-gram model is one of the most popular and easy forms of the language model which is suitable for long documents and short queries, it must apply a smoothing to eliminate zero counts. The Dirichlet smoothing model is the best smoothing model for short documents and short queries. Typically, the Dirichlet smoothing model uses a fixed value of the μ parameter which is equal to 2000 as mention in section 3.2. It has been adopted as the ideal value according to many empirical experiments. The μ parameter plays a strong role in finding the value of unseen terms as a contribution to avoid the zero-probability value. It is utilised to establish the amount of the mass of probability to be deducted from the viewed terms and to be added to the unnoticed terms, which depends on the length of the document and the mean probability of the viewed terms. The fixed value of the μ parameter becomes inappropriate and needs to be automatically estimated for structured documents [17]. In this research, it is not appropriate to predefine the μ parameter with a constant value and use it for different collections lengths. Among the estimation models, the likelihood model is the best estimation model for heterogeneous documents and short queries. In estimation model, the probabilities are calculated between query terms and document and then ranked the document based on their probabilities. DRO is a structured document where each document contains a large number of metadata units, these metadata units are contained in a single document containing different topics. In this case, if the estimation model done as usual the result will be entire documents which may be mostly irrelevant to the user query.

7. EXPERIMENTS AND RESULTS

The aim of this experiment is to provide an empirical justification to demonstrate that the LM model performance outperforms of other retrieval models such as vector space model and probabilistic model. CHiC2013 collection has been used as test collection. Table 5 shows some statistics related to CHiC2013 collection as well as the queries have been used. Furthermore, Table 6 presents the setting of LM regarding of DRO and based on the above observations. Table 7 reports the performance result for the CHiC2013 retrieving fewer than three models: LM, vector space model, and probabilistic model. The performance is measured by using the Mean Average Precision (MAP). From the table we can observe that the LM gets the 50% in term of MAP, while MAP value for the vector space model and probabilistic model are 29% and 39% respectively. It's clear that LM outperforms Vector Space Model and Probabilistic Model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5: Statistics of the test collection

Parameter Name	Value
Number of documents	1107
Number of metadata units in each document	1000
Number of testing queries based on documents retrieval	17

Table 6: Language model setting

N-gram	Unigram Model
Smoothing Model	Dirichlet Smoothing Model
Smoothing Parameter	μ =2000
Estimation Model	Query Likelihood Retrieval model $p(Q\|D)=\prod_{i=1}^{n} p(q_i\|D)$

Table 7: Comparison of MAP performance for the retrieving CHiC2013 based on: vector space model, probabilistic model and language model

Retrieval model	MAP
Vector Space Model	0.2933
Probabilistic Model	0.3902
Language Model	0.5013

8. CONCLUSION

In this research, a performance comparison among information retrieval models using a DRO (CHiC2013) collection has been presented. Further, an overview of the language model with its N-gram models, smoothing models, and estimation models have been presented. Moreover, comparisons in terms of advantage, disadvantage and usage for models have been given in different tables, and further studies related to language model have been summarized. Finally, various revised studies related to the language model have been systematically analyzed in terms of performance and the compatibility with DRO particularly in CHiC2013 collection, manifest that existing language models need to be adjusted before used in DRO to get along with its characteristics. for future works, the performance of DROs retrieval can be improved by enhancing parameter in the DS model to avoid the zero-probability value which leads to a decrease the DRO retrieval performance.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Universiti Sains Malaysia for this research under the Research University Grant entitled “A Two-Stage Expansion Method for Retrieval of Metadata Content in Cultural Heritage Collection”.

REFERENCES

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 5, pp. 179-190, 1983.

[2] Frederick, *Statistical methods for speech recognition*: MIT press, 1997.

[3] D. Hiemstra and F. De Jong, "Disambiguation strategies for cross-language information retrieval," in *International Conference on Theory and Practice of Digital Libraries*, 1999, pp. 274-293.

[4] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 275-281.

[5] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics,* vol. 19, pp. 263-311, 1993.

[6] S. Feng and R. Manmatha, "A discrete direct retrieval model for image and video retrieval," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, 2008, pp. 427-436.

[7] P. Ogilvie and J. Callan, "Combining document representations for known-item search," in *Proceedings of the 26th annual international ACM SIGIR conference on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Research and development in informaion retrieval, 2003, pp. 143-150.

[8] J. Xu, R. Weischedel, and C. Nguyen, "Evaluating a probabilistic model for cross-lingual information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 105-110.

[9] V. Lavrenko, M. Choquette, and W. B. Croft, "Cross-lingual relevance models," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 175-182.

[10] J. Xu and W. B. Croft, "Cluster-based language models for distributed retrieval," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 254-261.

[11] L. Si, R. Jin, J. Callan, and P. Ogilvie, "A language modeling framework for resource selection and results merging," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 391-397.

[12] G. Bennett, F. Scholer, and A. Uitdenbogerd, "A comparative study of probabilistic and language models for information retrieval," in *Proceedings of the nineteenth conference on Australasian database-Volume 75*, 2008, pp. 65-74.

[13] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval, Cambridge University Press. 2008," *Ch,* vol. 20, pp. 405-416, 2007.

[14] D. Jurafsky and J. H. Martin, *Speech and language processing* vol. 3: Pearson London, 2014.

[15] P. Banerjee and H. Han, "Language modeling approaches to information retrieval," 2009.

[16] C. D. Manning, C. D. Manning, and H. Schütze, *Foundations of statistical natural language processing*: MIT press, 1999.

[17] Y. Lv and C. Zhai, "Positional language models for information retrieval," presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, MA, USA, 2009.

[18] S. Petrovic, J. Snajder, B. D. Basic, and M. Kolar, "Comparison of collocation extraction measures for document indexing," *Journal of Computing and Information Technology,* vol. 14, pp. 321-327, 2006.

[19] J. Zhao and Y. Yun, "A proximity language model for information retrieval," presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, MA, USA, 2009.

[20] T. Westerveld and A. P. d. Vries, "Experimental result analysis for a generative probabilistic image retrieval model," presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.

[21] I. Vulić, W. De Smet, J. Tang, and M.-F. Moens, "Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications," *Information Processing & Management,* vol. 51, pp. 111-147, 2015.

[22] M.-H. Peetz, E. Meij, M. de Rijke, and W. Weerkamp, "Adaptive Temporal Query Modeling," in *Advances in Information Retrieval*, Berlin, Heidelberg, 2012, pp. 455-458.

[23] Y. X. Yu, "Personalization Information Retrieval Based on Unigram Language Model," in *Applied Mechanics and Materials*, 2013, pp. 2269-2273.

[24] S. Choi, J. Choi, S. Yoo, H. Kim, and Y. Lee, "Semantic concept-enriched dependence model for medical information retrieval," *Journal of biomedical informatics,* vol. 47, pp. 18-27, 2014.

[25] T. Baumel, R. Cohen, and M. Elhadad, "Topic Concentration in Query Focused Summarization Datasets," in *AAAI*, 2016, pp. 2573-2579.

[26] M. Symonds, P. Bruza, G. Zuccon, L. Sitbon, and I. Turner, *Is the unigram relevance model term independent? Classifying term dependencies in query expansion*, 2012.

[27] M. Jiang, E. Jensen, S. Beitzel, and S. Argamon, *Choosing the Right Bigrams for Information Retrieval*, 2004.

[28] R. Nallapati and J. Allan, "Capturing term dependencies using a language model based on sentence trees," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 383-390.

[29] K. Eguchi and W. B. Croft, "Query structuring and expansion with two-stage

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

term dependence for japanese web retrieval," *Information Retrieval,* vol. 12, pp. 251-274, 2009.

[30] A. Bruck and T. Tilahun, "Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique," *Journal of Computer and Communications,* vol. 03, pp. 67-76, 2015.

[31] M. Jiang, E. Jensen, S. Beitzel, and S. Argamon, *Effective use of phrases in language modeling to improve information retrieval*, 2018.

[32] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the eighth international conference on Information and knowledge management*, 1999, pp. 316-321.

[33] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language,* vol. 13, pp. 359-394, 1999.

[34] Zhai and Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems (TOIS),* vol. 22, pp. 179-214, 2004.

[35] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *ACM SIGIR Forum*, 2017, pp. 268-276.

[36] F. Jelinek and R. Mercer, *Interpolated estimation of Markov source parameters from sparse data*, 1980.

[37] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 111-119.

[38] Zhai, "Statistical Language Model for Information Retrieval," in *Tutorial Notes at the Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'06)*, 2006.

[39] G. Ding and B. Wang, "GJM-2: a special case of General Jelinek-Mercer smoothing method for language modeling approach to ad hoc IR," presented at the Proceedings of the Second Asia conference on Asia Information Retrieval Technology, Jeju Island, Korea, 2005.

[40] M. D. Smucker and J. Allan, "Lightening the load of document smoothing for better language modeling retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 699-700.

[41] D. E. Losada and L. Azzopardi, "An analysis on document length retrieval trends in language modeling smoothing," *Information Retrieval,* vol. 11, pp. 109-138, 2008.

[42] Zhai, Lafferty, and John, "Two-stage language models for information retrieval," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 49-56.

[43] Zhai, "Risk minimization and language modeling in text retrieval," PhD thesis, Carnegie Mellon University, 2002.

[44] He and Ounis, "A study of the dirichlet priors for term frequency normalisation," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 465-471.

[45] Zhai, "Statistical language models for information retrieval," *Synthesis Lectures on Human Language Technologies,* vol. 1, pp. 1-141, 2008.

[46] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language model-based search engine for complex queries," in *Proceedings of the International Conference on Intelligent Analysis*, 2005, pp. 2-6.

[47] L. Azzopardi and D. E. Losada, "Fairly retrieving documents of all lengths," in *Proceedings of the First International Conference in Theory of Information Retrieval (ICTIR 2007)*, 2007, pp. 65-76.

[48] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech & Language,* vol. 8, pp. 1-38, 1994.

[49] A. Berger and J. Lafferty, "Information retrieval as statistical translation," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 222-229.

[50] R. Nallapati, "Semantic language models for topic detection and tracking," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HLT-NAACL 2003 student research workshop-Volume 3, 2003, pp. 1-6.

[51] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of the tenth international conference on Information and knowledge management*, 2001, pp. 403-410.

[52] P. Bruza and D. Song, "A comparison of various approaches for using probabilistic dependencies in language modeling," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 419-420.

[53] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 171-180.

[54] Y. Lv and C. Zhai, "Positional language models for information retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 299-306.

[55] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 510-520.

[56] R. Cummins, J. H. Paik, and Y. Lv, "A Pólya urn document language model for improved information retrieval," *ACM Transactions on Information Systems (TOIS),* vol. 33, p. 21, 2015.

[57] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 1999, pp. 289-296.

[58] P. Banerjee and H. Han, "Answer credibility: A language modeling approach to answer validation," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009, pp. 157-160.

[59] B. J. Jansen, A. Spink, and J. Pedersen, "A temporal comparison of AltaVista Web searching," *Journal of the American Society for Information Science and Technology,* vol. 56, pp. 559-570, 2005.

[60] M. Spitters and W. Kraaij, "Using language models for tracking events of interest over time," in *In Proceedings of LMIR 2001*, 2001.

[61] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, USA, 2001.

[62] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, 2002.

[63] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," presented at the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom, 2004.

[64] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," in *Advances in Neural Information Processing Systems*, 2005, pp. 1617-1624.

[65] Tan, Z. Zhuang, P. Mitra, and C. L. Giles, "Efficiently detecting webpage updates using samples," presented at the Proceedings of the 7th international conference on Web engineering, Como, Italy, 2007.

[66] M. F. Caropreso, S. Matwin, and F. Sebastiani, "A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization," *Text databases and document management: Theory and practice,* vol. 5478, pp. 78-102, 2001.

[67] Tan, Y.-F. Wang, and C.-D. Lee, "The use of bigrams to enhance text categorization," *Information Processing & Management,* vol. 38, pp. 529-546, 2002/07/01/ 2002.

[68] A. Tombros, R. Villa, and C. J. Van Rijsbergen, "The effectiveness of query-specific hierarchic clustering in information retrieval," *Information processing & management,* vol. 38, pp. 559-582, 2002.

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[69] N. Bassiou and C. Kotropoulos, "Word Clustering using Long Distance Bigram Language Models," 2008.

[70] E. Dreyfuss, I. Goodfellow, and P. Baumstarck, "Clustering Methods for Improving Language Models," 2007.

[71] G. Erkan, "Language model-based document clustering using random walks," in *Proceedings of the main conference on human language technology conference of the north American chapter of the association of computational linguistics*, 2006, pp. 479-486.

[72] H. Mahmoodi and E. Mansoori, "Document Clustering Based On Semi-Supervised Term Clustering," *International Journal of Artificial Intelligence & Applications,* vol. 3, p. 69, 2012.

[73] S. Momtazi and D. Klakow, *A Word Clustering Approach for Language Model-based Sentence Retrieval in Question Answering Systems ABSTRACT*, 2009.

[74] M. A. Hearst, "Clustering versus faceted categories for information exploration," *Communications of the ACM,* vol. 49, pp. 59-61, 2006.

[75] K.-F. Wong, N.-K. Chan, and K.-L. Wong, "Improving document clustering by utilizing meta-data," in *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*, 2003, pp. 109-115.

[76] C. Laitang, K. Pinel-Sauvagnat, and M. Boughanem, "Estimating structural relevance of XML elements through language model," presented at the Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, Lisbon, Portugal, 2013.