EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Deep Level Markov Chain Model for Semantic Document Retrieval

Linh Bui Khanh[1], Ha Nguyen Thi Thu[1,*], Tinh Dao Thanh[2]

[1] {linhbk, hantt}@epu.edu.vn, Electric Power University, 235 Hoang Quoc Viet, Hanoi, Vietnam

[2] tinhdt@mta.edu.vn, Le Qui Don Technical University, 236 Hoang Quoc Viet, Hanoi, Vietnam

Abstract

The task of researching and developing information retrieval systems is becoming important in the big data age. Curren search methods try to mention to fast searching based on keyword matching or similar semantic between query anc documents but have not got a really effective engine for semantic search . In this paper, we propose a method fo information retrieval based on probability inference with the DLMC model to search by semantic equivalents and a topi word with score for fast searching. Results of the experimental with 952 Vietnamese documents show that our method is really effective for Vietnamese document retrieval system..

Keywords: Big data, information retrieval, feature reduction, Markov chain, probability inference.

Received on 29 April 2018, accepted on 18 July 2018, published on 10 September 2018

doi: 10.4108/eai.19-6-2018.155443

1. Introduction

Information on the Internet is created daily, hourly by millions of user on the global. Data on social networking site, personal blogs, personal website, businesses, etc., make online information more difficult to handle and can not to process with manual. Due to the increasing amount of information, to better store and process information, data mining tools need to propose and develop.

Information retrieval is a sub field of data mining with the goal of finding the information that matches the user's query. Information retrieval task are defined as aggregate tasks that include: representation, storage, organization, and access to information items. An information retrieval system described by the query processing process entered by the user from the interface. The query will be matched with the documents in the data store to return equivalent results, and then they are sorted in the same order as the query.

Due the important of the information retrieval task, a number of methods have been proposed [1-5]. Several models are used as language models, probability models, regression models; page rank and deep learning have proven their effectiveness in information retrieval [1-3], [5], [9]. Beside of this proposed model, some techniques for feature reduction are mentioned [8], [9], [11]. By reducing dimensional size, the number of features will be reduced, some features that are noise or non – important can be removed. It make information retrieval system more effectively, faster and more accuracy. Topic modelling is the one of model that used in some methods for feature reduction.

For some single syllable languages like Vietnamese, deny words does not based on space, so when working with Vietnamese language, people often use segmentation tools. Some current popular tools reach to over 90% accuracy. However, pre-processing and segmentation words in Vietnamese text will affect to speed of these systems. Thus, in the previous works [22], we refer to construction of notional word to expand topic modelling in order to better handle input queries. Because the topic modelling has a relatively small number of words, if the query does not contain the words in topic, it can be not to recognize documents that related with query. To overcome these difficulties we implement the following solutions:

- Built a notional words set to expand the search area of the query and reduce the pre-processing time of the query and documents in dataset.

*Corresponding author. Email: hantt@epu.edu.vn

- Use the deep level Markov chain model to infer similarity between the query and related documents.

The rest of the paper is structured as follows, in section 2, we present related work, in which we focus on two main issues of fast search and effective search with probability and language model. Section 3 proposed the Markov chain and improved Markov chain model. Deep level Markov chain and its application in information retrieval will be presented in Section 4. Experimental results of our method presented in section 5, finally is conclusion.

2. Related work

A number of studies have been proposed for increasing the efficiency of search engine, in which, some studies mentioned to indexing documents in database by organizing or placement methods. An others were concerned with the fast and accurate matching between queries and data sets.[4], [8], [16], [21].

There are many machine learning technical that proposed for information retrieval [1], [7], [15], [17], [19] but researchers always interested in feature selection and feature reduction because of its important [10-12]. The large dimensional data in real - life that make very hard to improve speed of calculating in any system. Furthermore, features in real – life sometimes make noise that affect to accurate of system. So that, there are a lot of studies concerned select and reduce feature problem. However, it is not easy to select feature and how to reduce it. In this paper, we select features (called notional words) include nouns, verbs and adjectives and maintain it in vocabulary of system.

Topic modelling is the basic model in natural language processing. Xing Wei [10] used topic modelling for information retrieval solution, they proved the effective of system when tested with TREC data sets after smoothing queries. Topic modelling was also proven effectiveness in narrow field information retrieval. Karla L Caballero Barajas used topic modelling in medical literature document retrieval [11]. Tuan Cao Xuan et., al also proposed a methods for mathematical document retrieval based on topic modelling [12].Ivan Vuli'c used topic modelling in cross – language information retrieval, they tested with three sets of data from CLEF 2001-2003 CLIR with a total of over 400,000 texts, compared to the BASELINE method with high accuracy [22]. We used own language model by creating a notional word model. This notional words set was built from Vietnamese dictionary combined with an extended words that include slang words, Internet term,etc.
Several probability models are still attractive many researchers when work with large data by learns from the data. Studies have shown that the accuracy of classical probability models may be better suited to identify similarities between query and document set or effective in ranking [6], [19]. A Markov chain model can infer from probability [15], [23], we used this model and improved it in depth level to better reflect on the information retrieval task.

3. Deep level Markov Chain Model

3.1 Markov Chain Model

The Markov model is one of the most important machine learning models in some task in NLP or some task in predicting [23], [24]. To define it properly, we need to first introduce the Markov chain, sometimes called the observed Markov model. Markov chains is extensions of the finite automat. Recall that a weighted finite automaton is defined by a set of states and a set of Markov chain transitions between states, with each arc associated with a weight. A Markov chain is a special case of a weighted automaton in which weights are probabilities (the probabilities on all arcs leaving a node must sum to 1) and in which the input sequence uniquely determines which states the automaton will go through. Because it can't represent inherently ambiguous problems, a Markov chain is only useful for assigning probabilities to unambiguous sequences [23], [24].

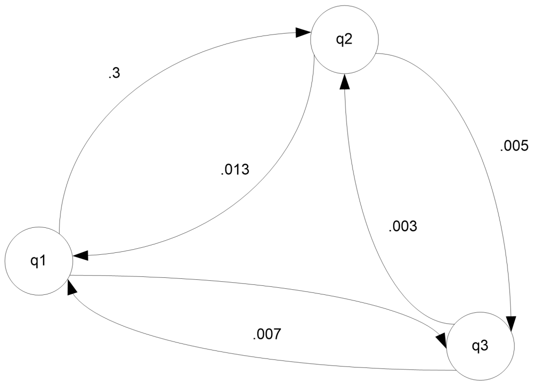

Figure 1. Markov chain

In figure 1 illustrated a Markov chain that includes three states q_1, q_2 and q_3. These arcs from q_i to q_i are called

transition probability from states.
In formal, A Markov chain includes parameters:

$Q = q_1 q_2 .. q_N$	A set of N states
A=a01a02…an1…anm	A transition probability matrix A, each aij representing the probability of moving from state I to state j. Total probability from a=1
Q0,qF	Start state and end state

A Markov chain embodies an important assumption about these probabilities. In a first-order Markov chain, the probability of a particular state depends only on the previous state:

$$P(q_i \mid q_1..q_{i-1}) = P(q_i \mid q_{i-1}) \qquad (1)$$

3.2 Deep Level Markov chain model

A deep level Markov chain model that improved based basic Markov chain model. In which, we implement more parameter ϑ for each state, however this state is not observation state that inferred from trained set (So that, it is not a hidden Markov model).
A deep level Markov chain model contains three parameters:

Q= q1q2..qN	A set of N states
$A = a_{01}a_{02}...a_{n1}...a_{nm}$	A transition probability matrix A, each a_{ij} representing the probability of moving from state i to state j. Total probability from a=1
$I = i_1,..,i_k$	Observation probability set of each state

In which, number of the observers is equal to each observation.

4. Methodology of Information retrieval

4.1 Notional words set

Notional words set include nouns, adjectives and verbs. In formal notional word set has been defined as follows

$$\Gamma = \{\tau_i \mid \tau_i \in N, V, ADJ\} \qquad (2)$$

In which:

- N: noun
- V: verb
- ADJ: Adjective

A notional word set has been built in two ways. The first, we extracted from Vietnamese lexicon, however sometimes lexicon does not cover the wide range of words on the Internet, acronyms and slang. So that, we extended this notional word set from a document set. These documents were downloaded from the Internet, after that, notional words was been extracted from its.

4.2 DLMC – Information retrieval Model

Set $q = \{q_1, q_2,.., q_n\}$ is the notional words that extracted from query. D is the document set and denoted by $D = \{d_1, d_2,..., d_n\}$. In each document d_i in D, we can extract notional word set $d_i = \{w_{i1}, w_{i2},.., w_{in}\}$.

The task of determining a keyword q_k is contained in query or not by matching often brings rapid results. In cases, semantic of query is equivalent with the document but they have not any the same terms. Therefore, the search engine can miss this query. To overcome this, we used a Deep level Markov chain model (Fig 3) for inferring semantics. In which, query q has been separated by notional words $q_1, q_2,.., q_n$ and they are considered the states of DLMC model. The transactions between states are illustrated by arcs, but in this model, no arrows on the arcs and it also have no start and end state. There are no direction movement on the arcs. On each state, we can infer a set of document D and use similarity of each state with each document d_i in D for retrieving.

Definition 1. Topic related score

Topic related score is defined the co – occurrences of terms in a document. It is calculated by

$$C(\mathrm{w}_1, \mathrm{w}_2) = \frac{\text{number of document that occurrence } \mathrm{w}_1 \text{ and } \mathrm{w}_2}{\text{total of document in training set}}$$

(3)

Definition 2. Compatible score

Compatible score express document d is satisfied with query Q and it is calculated by:

$$\xi(Q,d) = \alpha \sum_{i=1,m=1}^{x,y} \mathrm{C}(\mathrm{q}_i, \mathrm{w}_m) + \beta \sum_{i=1,j=i+1}^{x} C(q_i, \mathrm{q}_j)$$

(4)

In which:

- $\mathrm{C}(\mathrm{q}_i, \mathrm{w}_m)$: Topic related score between each notional words in query Q and document d. $C(q_i, \mathrm{q}_j)$: Topic related score between qi in query Q.
- α , β : Coefficients

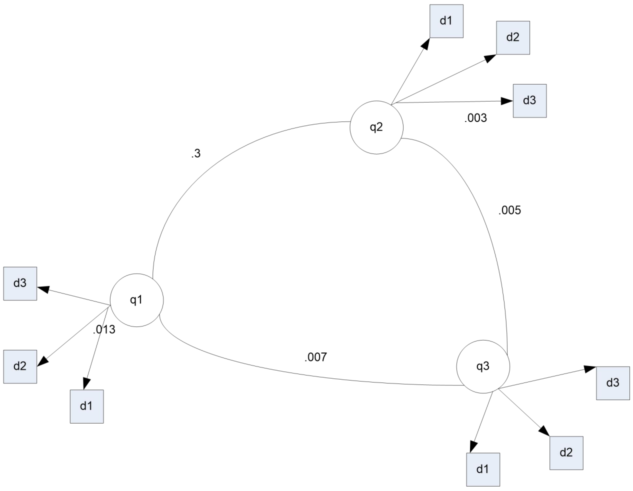

Figure 2. DLMC Model for Information retrieval

Figure 2 describes the DLMC model. Given the query Q is separated into 03 notional words q_1, q_2 and q_3. These notional words are interconnected with each other by weighted arcs. Each notional word q_j in query Q is matched with document d_i in document set D for finding the relation between notional word qj and document d_i. Weighted in associated arcs of notional words $arc(q_1,q_2), arc(q_2,q_3)$ and $arc(q_1,q_3)$ is calculated by the topic related score by formula (3). The relation of query Q and document d_i is calculated by formula (4).

The algorithm for information retrieval is presented by follows:

```
INFORMATION RETRIEVAL BASED ON DLMC
Input
    Q: query, $\Gamma$ : notional word dictionary, D: set of documents
Output:
    $\Upsilon$ : Set of documents that related with query Q
1.1 Select feature of query Q
    For i=1 to length(Q) do
    If q[i] $\in \Gamma$ then
    select them as feature of query Q
1.2 Select feature of each document in D
    For each document $d_k$ in D
    For i=1 to length($d_k$) do
    If $w_k$[i] $\in \Gamma$ then
    select them as feature of document $d_k$
1.3 Calculate compatible score between query Q and document
    Calculate topic related score between notional word in query Q
    Calculate topic related score between notional word in query Q and document d
    Compatible score by
    $$\xi(Q,d)=\alpha\sum_{i=1,m=1}^{x,y} \mathrm{C}(\mathrm{q}_i,\mathrm{w}_m)+\beta\sum_{i=1,j=i+1}^{x} C(q_i,\mathrm{q}_j)$$
```

5. Experimental

5.1 Notional word set

We used Vietnamese document for experimental. Due to the lack of corpus in Vietnamese document retrieval field, so that some researchers still use self- collected data. In order to obtain empirical data, we download from news pages like http://vietnamnet.vn, http://vnexpress.net, and pre-processing. This data used for training, building notional word set. There are 952 documents in this corpus.

A part of notional words we extracted from Vietnamese dictionary like table 1 bellows, but in the corpus we have 1,022 words.

Table1. Some notional words extracted from dictionary

word	pos	
a dua	đg.	
a hoàn	d.	
a phiến	d.	
a tòng	đg.	
à ơi	c.	
ả		2
ả đào	d.	
á nguyên	d.	
á phiện	d.	
ác		3
ác	t.	1
ác		3
ác báo	đg.	
ác cái là		
ác giả ác báo		
ác hại	t.	
ác thần	d.	
ác thú	d.	

For addition to this dictionary by slang words, Internet terms, we extracted notional words from the corpus. Here is some words that are extracted from document set D. For this corpus, we extracted 234 notional words.

Table 2. Some notional words extracted from corpus

word	pos	
Selfie	đg	
wall		
add	Đg	
Chat		
comment	Đg	
tag	Đg	2
HF	d.	Hot face

After completing the notional words, we assigned score by formula (3). In table 3, we illustrated some pair of notional words with their co – occurrences score.

Table 3. Co - occurrences words with weighted

	Nói (talk)	Hoa (flower)	Chân (foot)	Trắng (white)
Cảnh (sight)	0.00520833	0.19965278	0.02256944	0.00347222
Bình yên (quiet)	0.01128472	0.03993056	0.01215278	0.10763889
Thắng lợi (success)	0.05208333	0.3671875	0.13541667	0.00260417
Mây (cloud)	0.00130208	0.27083333	0.0078125	0.27864583
Chuối (banana)	0.00086806	0.39583333	0.02951389	0.33072917
bàn (table)	0.00173611	0.04861111	0.00217014	0.00173611
Chuẩn bị (prepare)	0.21267361	0.02951389	0.02083333	0.00434028
Nhảy (jump)	0.20399306	0.00303819	0.30902778	0.00607639

5.2. Evaluating

At the present time, there is no automated tool for evaluating Vietnamese document retrieval. Therefore, we used precision and recall for evaluation. In order to evaluate, we built a system that based on C# language and SQL Server.

$$precision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN}$$

In which:

- TP – true positive: These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes.
- FP - false positive: When actual class is no and predicted class is yes
- FN - false negative: When actual class is yes but predicted class in no
- TN - true negative: here are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

Table 4 is the results of this methodology

Table 4 Precision and recall

Topic	Rec	Prec
Business	0.373	0.894
Sport	0.156	0.951
Technology	0.191	0.929
Education	0.391	0.868
Travel	0.356	0.878
Politics	0.267	0.93
Financial	0.42	0.78

We used average of results for comparing with an others system. In this paper two systems are used for evaluating with Lucence system [13] and a method in [12] shown that: Precision of our method is 0.862069 and [12] has precision 0.816092 with the same corpus.

- The system in [12] only calculates the similarity between the query and clusters, it does not mention to rank documents by relation score. Therefore, there are a number of documents can be retrieved but it is not related to query.

- The system in [12] only recognizes well when the documents contain mathematical formulas and topic modeling in narrow field.

- The system in [12] sometimes cannot find documents that related to query when query does not contains terms that contains in topic modeling.

6. Conclusion

With the rapid growth of information on the Internet, require more and more tools for processing and mining data. So that needs to propose methods or solutions for improving information retrieval system.

In this paper, we proposed a method that used linguistic model and Markov chain to speed up the information retrieval system based on notional words and its co – occurrences score. Firstly, we do not have to integrate Vietnamese processing tools for processing queries and documents. After that, we improved the Markov chain model at deeper level for ranking document that related to query. So that, our method can reduces processing time, increase speed of system with high accuracy.

References

[1] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In Proceedings of the 20th Australasian Document Computing Symposium, pages Article–No. ACM, 2015.

[2] Vuli´c and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 363–372. ACM, 2015.

[3] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 795–798. ACM, 2015.

[4] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1188–1196, 2014.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. CJorrado, and M. I. Dean, Jeffdan. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[6] Norbert Fuhr. A probability ranking principle for interactive information retrieval. Information Retrieval, 11(3):251–265, 2008.

[7] Kirsty Kitto, Peter Bruza, Liane Gabora, A Quantum Information Retrieval Approach to Memory, Proceedings of the International Joint Conference on Neural Networks, (pp. 932-939). June 10-15, Brisbane, Australia, IEEE Computational Intelligence Society.

[8] Ha Nguyen Thi Thu, Linh Bui Khanh, Tinh Dao Thanh and Vinh Ho Ngoc, An Effective Data Organizing Method for Vietnamese Document Retrieval, ICIC Express Letters, An International Journal of Research and Surveys, pp.955=960, Vol. 11, No.5, 2017.

[9] Shariq Bashir, An Improved Retrievability-Based Cluster-Resampling Approach for Pseudo Relevance Feedback, Computers 2016, 5, 29; doi:10.3390/computers5040029.

[10] Xing Wei, Topic Models in Information Retrieval, Ph.D. dissertation, University of Massachusetts, Amherst, MA, 2007.

[11] Karla L. Caballero Barajas, Ram Akella:Incorporating Statistical Topic Models in the Retrieval of Healthcare Documents. CLEF (Working Notes) 2013.

[12] Tuan Cao Xuan, Linh Bui Khanh, Hung Vo Trung, Ha Nguyen Thi Thu, Tinh Dao Thanh “Indexing Based on Topic Modeling and MATHML for Building Vietnamese Technical Document Retrieval Effectively.” ICCASA 2015: 322-332.

[13] https://vlsp.hpda.vn/demo/?&lang=en

[14] https://lucene.apache.org/

[15] Hanjun Dai, Bo Dai, Yan-Ming Zhang, Shuang Li1, Le Song, Recurrent Hidden semi – Markov Model, ICLR 2017 conference.

[16] Do Thi Thanh Tuyen, Nguyen Tuan Dang, Phrasal Semantic Distance for Vietnamese textual Document Retrieval, Journal of Computer Science and Cybernetics, V.31, N.3 (2015), 185 202.

[17] Peter Nabende, Jörg Tiedemann, John Nerbonne, Pair Hidden Markov Model for Named Entity

[18] Matching, Innovations and Advances in Computer Sciences and Engineering pp 497-502.

[19] David R. H. Miller, Tim Leek, Richard M. Schwartz, A Hidden Markov Model, Information Retrieval System, SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval pp. 214-221.

[20] J. Szyma_nski and W. Duch, _Information retrieval with semantic memory model,_ Cognitive Systems Research, vol. 14, no. 1, pp. 84_100, 2012.

[21] Khanh Linh Bui, Thi Ngoc Tu Nguyen, Thi Thu Ha Nguyen and Thanh Tinh Dao, An Effective of Data Organizing Method Combines with Naïve Bayes for Vietnamese Document Retrieval, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 205-219, 2017.

[22] Ivan Vuli´c, Wim De Smet, and Marie-Francine Moens, Cross-Language Information Retrieval with Latent Topic Models Trained on a Comparable Corpus.

[23] Faten Khalil, Jiuyong Li and Hua Wang, A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses, Proc. Fifth Australasian Data Mining Conference (AusDM2006), pp. 177-184.

[24] F Khalil, H Wang, J Li, Integrating Markov Model with Clustering for Predicting Web Page Accesses, Proceeding of the 13th Australasian World Wide Web Conference (AusWeb07), 63-74, 2007.