DESIDOC Journal of Library & Information Technology, Vol. 31, No. 4, July 2011, pp. 242-246

Implicit Semantic Relations Identification through Distributed Representations for Effective Text Retrieval

Rajendra Prasath

Indian Institute of Technology, Kharagpur-721 302,
E-mail: rajendra@cse.iitkgp.ernet.in; drrprasath@gmail.com

ABSTRACT

The modern commercial search methods have proved the efficiency of information retrieval (IR) technologies and made knowledge dissemination simpler where as finding the relevant text documents, given a query, is becoming much more complex day after day. Almost all information retrieval systems struggle to retrieve relevant entries on the top of ranked list of documents. Users are interested to retrieve the content satisfying their information needs rather than just retrieving the documents having the query terms supplied by them. For this task, traditional information retrieval methods hardly yield the results having higher-order relations between the given query and the documents in the collection. So, it becomes necessary to find and encode the higher-order feature associations present in text documents. This paper discusses the use of a distributed representation, namely random indexing for an effective retrieval of relevant text documents. This type of distributed representation would be scalable with modern computing facilities and flexible to develop knowledge-based applications, which may require the process of identifying implicit semantic relations though higher-order term associations.

Keywords: Distributed representations, dimensionality reduction, random indexing, holographic reduced representations, higher-level relations, knowledge-based systems

1. INTRODUCTION

Different types of digital contents are growing rapidly with the advancements of World Wide Web (WWW) technologies like internet, e-mail, news service, blogging, messaging like tweek, buzz, and so on. The organisation of such huge volumes of digitised information is becoming a complex task and needs human expertise to put them under proper tagging/ representation so as to retrieve them at a later stage, possibly with little effort. Even though efficient indexing strategies have been applied to organise the content, especially the textual content extracted from the digital collection, still, searching and retrieving the relevant information from such huge collection satisfying the information needs of the users is a highly challenging task. Modern search engines keep on applying improvements to their ranking algorithms so as to retrieve the most relevant documents that meet the information needs of the users.

On the other hand, users provide a set of keywords to express their information needs. Most often, these keywords are hardly sufficient to understand the information needs of the users behind the supplied query. So, methods to efficiently encode the associations among the features[1] or terms in the documents, and then the terms in the query, are needed so that similarity estimation to retrieve the most relevant documents would be much easier. Another major challenge in information retrieval (IR) systems is to deal with incomplete or underspecified information in the queries submitted by the users. The IR systems receiving such queries need to fill the gaps in the underspecified queries of users so as to improve the retrieval efficiency[2].

The problems multiply in cross-lingual domains where cross-lingual information retrieval (CLIR) systems often use translation to cross the language barriers between a query and the relevant documents. However, the query translation and translation disambiguation of underspecified queries make the retrieval and ranking of documents more challenging in CLIR. In this scenario, it is important to study more appropriate ranking algorithms for CLIR systems.

Pirkola proposed a query structuring method for grouping query keywords. This method further suggests the use of query operators in such a way that more weight is being assigned to important or correct keywords than the other keywords[3]. This query structuring captures the clue on the intention of users' information needs with bilingual English-Finnish news data. However, this method does not capture semantic associations among query terms and the terms in the documents. In this attempt, mechanisms have been presented that represent documents in an encoded form. Then retrieval task is performed by identifying the query type that gives a clue on the information needs of the users across multiple language contents and then to rank these retrieved documents in a better order. The methods explained in this paper may be tested against the standard benchmark IR test collections provided by TREC, CLEF and FIRE and standard TREC evaluation metrics may be used to measure the retrieval efficiency[4].

2. PAST WORK

Data mining and knowledge discovery forced the people to think about how to use natural language processing and information extraction methods to automatically extract relevant factual information Department of Computer Science and Engineering? Here, retrieval was not only the focused task but reuse of textual description in free text form without converting it into structured cases is also of special interest. To make the reuse of textual descriptions better, it was essential to understand the meanings of query terms along with their contexts/relations. Two types of relations would be essential to investigate the sequel: (i) associate relations: immediate relations with adjacent terms like eat -> food, and (ii) synonymy relations: higher-order relations of the focused terms that share their contexts like eat -> drink. Word space methodology helped to understand such relations. These relations would make semantics computable and would constitute a purely descriptive approach to semantic modelling. Additionally, words with similar meanings would tend to occur in similar contexts. Distributed representations extract semantically similar words and capture their contexts as well. Many algorithms have recently been developed using distributional representations to capture such semantic relations. Motivated by such distributional hypothesis, random indexing, which efficiently captures higher-order term associations in text descriptions, has been illustrated.

3. DISTRIBUTED REPRESENTATIONS—SEMANTIC RELATIONS

With the invention of latent semantic indexing (LSI), the exploration of semantic structures began. LSI takes the advantage of implicit higher-order structures associating terms and documents to improve the retrieval of relevant documents (Deerwester). LSI requires processing of the huge, sparse term—document matrix and singular value decomposition (SVD) is applied on that matrix to get the reduced set of orthogonal factors from which the original matrix can be approximated by linear combination. Thus, LSI first constructs a huge term, document matrix, and then uses a separate dimension-reduction phase by applying SVD. This method has many inherited disadvantages: Whenever a new update to the feature set is made, SVD has to be re-applied on the newly updated matrix, which is computationally expensive. This method is not incremental as newly added feature may result in modified term document matrix. Random indexing (RI), a fine alternative to LSI, is an incremental word space model based on sparse distributed representations. The main idea behind RI is to accumulate term context vectors based on the occurrence of words in specific contexts.

3.1 Random Indexing

The main intuition behind RI was derived as a result of Johnson-Lindenstrauss Lemma[5]. It suggests the fact that whenever a set of points in a higher dimensional is mapped into a reduced dimensional space, the distance between any pair of points does not vary significantly[6-10]. This lemma was used for the computational procedure behind random indexing in the sequel. Here, two types of feature vectors have been used: Feature Index Vectors and Feature Context Vectors.

3.1.1 Feature Index Vectors

Each feature (may be a word or a document) is assigned a unique and randomly generated representation of fixed-size vectors. Index vectors are sparse, high-dimensional, and ternary and their dimensionality (n) is in the order of thousands. Also, they comprise a small number of randomly distributed +1s and -1s, with the rest of the elements of the vectors set to 0.

3.1.2 Feature Context Vectors

Each context vector is produced by scanning through index vectors of the feature present in the given text fragment. Whenever a feature occurs in the context within a sliding context window of fixed size, the context vector of that feature is updated by adding, and its n-dimensional index vector multiplies with real coefficient, which is based on the position of the occurrence of the term. Features are thus represented by n-dimensional context vectors that collectively represent the sum of the features' contexts. For each occurrence of a given feature, focus is made on a fixed window of size $(2 \times k) + 1$ centered at the given feature [suggested window size is 5 (terms + / - k

terms; here $k = 2$)]. The feature context vector for feature$_i$ is then computed using the following equation:

$$C_{feature_i} = C_{feature_i} + \sum_{j=-k; j\neq 0}^{+k} I_{feature(i+j)} \times \frac{1}{d^{|j|}}$$

where $d^{|j|}$ is the weight proportion wrt the size j of window ($d = 2$). Superposition is used to add two context vectors during the training of feature context vectors.

4. EVALUATION METHODOLOGY

Document weight is computed as:

$$\frac{w_{c_i}}{\sqrt{\sum_{i=1}^{m} w_{c_i}^2}}$$

where w_{c_i} is the superposition of context vectors of the features in the document, each multiplied with its frequency.

Query weight is computed as:

$$\frac{w_{q_i}}{\sqrt{\sum_{i=1}^{m} w_{q_i}^2}}$$

where w_{q_i} is the superposition of context vectors, of the query terms, each multiplied with its frequency.

The cosine similarity is used to retrieve the ranked list of documents matching higher-order term associations following:

$$\text{sim}(q_i, c_i) = \sum_{\text{matching features}} W_{q_i} \times W_{c_i}$$

The effects of document retrieval using RI can be explained with a simple example. Consider the following six documents each having two features:

(i) c1: *feature*$_A$ and *feature*$_B$;

(ii) c2: *feature*$_B$ and *feature*$_C$;

(iii) c3: *feature*$_C$ and *feature*$_D$;

(iv) c4: *feature*$_D$ and *feature*$_E$;

(v) c5: *feature*$_E$ and *feature*$_F$;

(vi) c6: *feature*$_F$ and *feature*$_G$.

The content of the documents c1 and c2 are taken as query terms separately. The retrieved ranked list is shown in Table 1. The standard term Frequency—Inverse Document Frequency (TF-IDF) is used as the baseline method and random indexing results are compared with it as tabulated in Table 1.

Table 1. Retrieval efficiency by TF-IDF and RI

New case	TF-IDF		Random Indexing	
	Best hypotheses	Similarity	Best hypotheses	Similarity
c1	c1	1.0000	c1	1.0000
	c2	0.3696	c2	0.8698
	c3	0.0000	c3	0.5826
	c4	0.0000	c4	0.2865
	c5	0.0000	c5	0.0095
	c6	0.0000	c6	0.0055
c2	c2	1.0000	c2	1.0000
	c3	0.5000	c1	0.8698
	c1	0.3696	c3	0.7636
	c4	0.0000	c4	0.5145
	c5	0.0000	c5	0.2603
	c6	0.0000	c6	0.0142

Random indexing performs dimension reduction implicitly in an effective way and dimensionality of the vectors is a parameter. RI is incremental at runtime and provides ways not to affect the entire collection of trained context vectors. Only relevant context vectors of the features have been fetched, modified and updated with the newly seen contexts. Additionally, it does not requires a separate dimension-reduction phase.

5. APPLICATIONS

The following are the interesting applications having the potential to make use of distributed representations are:

5.1 Digital Libraries

Modern digital library systems use sophisticated search facilities. Several abstracting services are quite popular and all these services require the involvement of humans in the subject/topic classification. This task is laborious and requires in-depth domain knowledge to classify the documents under the categories of that domain.

If a person, new to research, is interested in exploring particular issues of a problem with its related analogies across multiple domains through the abstracting services, then he/she hardly gets a clue other than searching with appropriate keywords. In such cases, to make the search intelligent, applications using

distributed representations may be used to retrieve semantically-related analogies of a given problem or the applications of a particular theory.

5.2 Cross-Lingual Information Retrieval

Multilingual content in the WWW is growing faster as users prefer to share their knowledge in their own language. Like India and Europe, there are many countries having more than one language and Web content is growing in the native language of users. With the growth of vast amount of such multilingual Web content, it becomes essential for users to access the information present across different languages. Research on cross-lingual information access has emerged to assist the information needs of the community with different language speaking users who may issue queries in one language by enabling them to access the information written in other languages[11,12]. CLIR systems consists of many complex tasks like query translation/ transliteration, query (re)formulation, word-sense disambiguation, identifying semantically equivalent term variations during query expansion, and so on. Distributed representations can be suitably applied to these tasks so as to identify semantically associated documents retrieval rather than just applying "bag of words" approach or vector space model[13].

5.3 Textual Case-based Reasoning

Case-based reasoning (CBR) is emerging as an important sub-topic of artificial intelligence. The main idea of CBR is how to automatically use past user experiences to solve similar future problems. Here, the term "similar" is loosely used and rather "semantically similar" would suit the context better. In classical CBR, given a case base having structured cases of the past experience and a new situation (new case), extracting the most relevant cases similar to the new situation is not just matching the important attribute–value pairs identified in the new situation. But the similarity assessment between the new situation and the case retrieved from case base should account the higher-order term relationships, particularly in the models of analogical reasoning[1].

Whenever the past experience is available in the form of natural language text documents like reports, monographs, etc., this process becomes more complex; the problem of identifying the informative features reduces to the problem of identifying certain rhetoric structures having discriminative features like attribute – value pairs in the given textual descriptions. Here the values of the features increase with the number of its occurrences in the case collection. In such situations, one could apply document length normalisation. In the mean time, any normalisation factor has an effect of decreasing the weight of the document features thereby reducing the chances of retrieval of the document. Therefore, the higher the normalisation factor for a document, the lower is the chances of retrieval of that document[14]. Domains requiring such types of investigations include oil well drilling, accidental reports, financial reports/forecasting, monsoon reports, disaster reports, safety reports, complaint registers, customer opinions, and so on.

5.4 Text Classification

Text categorisation is an essential task in organising the documents with proper tagging on its types. This assists the searching faster by making the search domain focused on a sub-set of documents with matching type as of the given query. Recently, semantic analysis techniques have largely been applied to arrive at higher speed and accuracy in classification. External knowledge sources like Wikipedia, ODP, etc. are used as a collection of knowledge concepts from which the meaning of words and text segments are interpreted. Still this process needs the high cost of calculation, language, and bandwidth to access and process the knowledge resources and is not cost-effective for research purpose. Recently, attempts have been made to use distributional features for text categorisation. Using distributional feature representations, different features can differently be weighted based on their importance in the given context. In web context, different weighting functions may be used to assign different weights according to their position[15].

6. CONCLUSION

The proposed way of encoding/presentation for textual content enables the dimension reduction in an effective way, doing it implicitly, not as a separate stage, as in latent semantic indexing. This method takes into consideration both semantic and structural properties of features. Additionally, these compressed representations allow learning the meaning of features and documents through experiences incrementally and in depth, whereas the similarity assessment becomes affordable without heavy computations.

REFERENCES

1. Gentner, D. & Forbus, K.D. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 1991, **19,** 141-205.

2. He, D. & Wu, D. Enhancing query translation with relevance feedback in translingual information retrieval. *Inf. Process. Manage.*, 2010, **47**(1), 1-17.

3. Pirkola, A.; Puolamaki, D. & Jarvelin, K, Applying query structuring in cross-language retrieval. *Inf. Process. Manage.*, 2003, **39**(3), 391-402.

4. Sanderson. M. Test collection-based evaluation of information retrieval systems. *Foundations Trends Inf. Retri.*, 2010, **4**(4) 247-375.

5. Johnson, W. & Lindenstrauss, L. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics,* 1984, **26**, 189-206.

6. Kanerva, P.; Kristofersson, J. & Holst, A. Random indexing of text samples for latent semantic analysis. *In* Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Erlbaum, Mahwah, 2000. pp. 103-06.

7. Öztürk, P. & Prasath, R. Recognition of higher-order relations among features in textual cases using random indexing, *In* Proceedings of 18th International Conference on Case-Based Reasoning. *Lecture Notes Compu. Sci.*, 2010, **6176**, 272-86.

8. Öztürk, P., Prasath, R. & Moen, H. Distributed representations to detect higher order term correlations in textual content, *In* Proceedings of 7th International Conference on Rough Sets and Current Trends in Computing. *Lecture Notes Compu. Sci.*, 2010, **6086,** 740-50.

9. Sahlgren, M. Vector-based semantic analysis: Representing word meanings based on random labels: *In* ESSLI Workshop on Semantic Knowledge Acquisition and Categorization. Kluwer Academic Publishers, Dordrecht, 2001.

10. Sahlgren, M. An introduction to random indexing. Methods and Applications of Semantic Indexing Workshop. *In* Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, 2005.

11. Majumder, P.; Mitra, M. & Datta, K. Multilingual information access: An Indian language perspective *In* Proceedings ACM SIGIR Workshop on New Directions in Multilingual Information Access, Seattle, 2006.

12. Prasath, R. & Öztürk. P. Similarity assessment through blocking and affordance assignment in textual CBR. *In* Proceedings of Reasoning from Experiences on the Web, 2010. pp.151-60.

13. Salton, G.; Wong, A. & Yang, C.S. A vector space model for automatic indexing. *Commnication ACM,* 1975, **18**(11), 613-20.

14. Singhal, A.; Salton, G.; Mitra, M. & Buckley, C. Document length normalisation. *Inf. Process. Manage.,* 1996, **32**(5) 619–33.

15. Xue, X.B. & Zhou, Z.H. Distributional features for text categorisation. *IEEE Trans. Knowl. Data Engg.*, 2009, **21**(3), 428-42.

About the Author

Dr Rajendra Prasath did his MSc (Mathematics) from Rumanian Institute for Advanced Study, MTech (CSE) from Indian Institute of Technology, Kharagpur, and PhD (Mathematics-Computer Science) from University of Madras, Chennai. He started his research career with a guest faculty position at University of Madras in 1998. During 2004-2006, he worked as Assistant Professor at MNMJEC under Anna University, Chennai. Later he joined Communication Empowerment Laboratory of IIT Kharagpur as a Senior Project Officer. During August 2009 to September 2010, he visited the Norwegian University of Science and Technology (NTNU), Norway, as an ERCIM Alain Bensoussan Fellow. Rajendra was a Visiting Fellow to the Artificial Intelligence Research Institute (IIIA); Spanish National Research Council (CSIC), Barcelona, Spain, and Swedish Institute of Computer Science (SICS), Kista, Sweden during May-June 2010. Earlier, he was a University Research Fellow at University of Madras during November 2001 to April 2003. He contributed tools to Cross-Lingual Information Access system (at IIT KGP) which was a part of DIT, Govt. of India sponsored research work. Presently he is a Technical Editor for the *Advances in Information Sciences* and *Journal of Computer Science*. He has also served as a reviewer for *IEEE/ACM Transactions on Networking, Information Sciences, Journal of Convergence Information Technology* and several international conferences. He is a member of World Federation on Soft Computing, Information Retrieval Facility, Vienna, International Rough Set Society (IRSS), Warsaw, and Information Retrieval Society of India. His research interests include cross-lingual information retrieval, textual case-based reasoning, machine learning, and distributed algorithms for message passing systems.