An Information Retrieval Expansion Model Based on Quasi-Clique

Lixin Gan
School of Math and Computer Science, Jiangxi Science & Technology Normal University
Nanchang, China
spiderganxin @163.com

Wei Tu
Center of Arts Complex Lab, Jiangxi Science and Technology Normal University
Nanchang, China
ncsytuwei @163.com

Ying Xiong
Jiangxi Science & Technology Normal University
Nanchang, China
2378008613@163.com

***Abstract*—Query expansion is an important technology for improving retrieval performance in information retrieval. Many Studies have found contexts within query that strongly influence the interpretation of a query. In this paper, we propose the graph mining technique called Quasi-Clique as query context in Markov network retrieval model. Our approach exploits contextual information mined from the term Markov network for per query term in addition to syntactic similarity. The proposal in our work is benefit to select more relevant terms for query expansion and to improve precision in information retrieval.**

Keywords—information retrieval; query expansion; quasi-clique; Markov network retrieval

I. INTRODUCTION

With the rapid development of Internet, our world has come into big data era. How to get exact information rapidly and expediently is becoming a problem which is needed to be solved urgently. Despite the recent advances in search engineer, big data has introduced a great of new challenges. Therefore, query expansion, as one of the key technologies to improve recall and precision in information retrieval, has been a strong research interest [1-3]. Typical approaches use one or a combination of sources to generate additional query terms.

By and large, most of previous approaches to the query expansion problem only consider the simple dependence between terms. If the similarity is more than certain threshold, then the pair of terms becomes candidates for query expansion. These approaches often suffer from a large number of false positives so that the topic drifts and the overall performance suffers. The essence of this false positive problem is that these approaches take use of use of distance functions that solely rely on the "textual similarity" of two terms, regardless of the adopted graph structures. In reality, terms, which are strongly similar to the query topic rather than a single query term, are more important and benefit to improve efficiency. Such kind of terms should be added to original query. Therefore, a method called context often shows a good performance if it considers some "additional information" beyond textual similarities. Contextual information is captured as context graph.

Toward this problem, some models have been proposed based on context concept, such as Semantic Relationship Graph [4]. Paper [5] proposes a latent concept expansion which uses clique for modeling term dependencies during expansion. However, this model yields no conclusive results with regard to expansion using multi-term concepts. In paper [6], an information retrieval model based on Markov concept is proposed which learns from document set, and shapes of concept mined from Markov Network such as clique and Markov concept graph are added into query process. The models proposed in [6] perform outstanding and make significant improvements. A novel query expansion technique based on concept clique for Markov network information retrieval model was proposed in [7], which integrates query term dependency to the selection of concept clique. There is a hypothesis in [7] that if query terms within a query have relationships, their concept cliques as candidates will be in a connected graph. Therefore, when expanding dependent query term, although some single concept cliques with high similarity to original query terms, they will be considered as noise and be not taken into account. Cliques is uses to as query context which refers to context words within the query in[8] and the approach is particularly useful for the selection of relevant term relations to avoid topic drift. However, a clique mentioned above in previous work requires that every two terms are directly connected with each other. However, the conditional of a clique is so strict that some terms with high similarity to some original query term but not in a clique will be not taken into account for query expansion. To solve this problem, in this paper, we propose the graph mining technique, Quasi-Clique, as query context in Markov network retrieval model. Our approach exploits contextual information mined from the network of terms for per query term in addition to syntactic similarity. The proposal in our work is benefit to select more relevant terms for query expansion, to reduce topic drift and to improve retrieval performance.

II. RELATED WORK

Markov network is an undirected graphical network that is capable of efficiently representing relevance in knowledge and can be easily gotten from training data with strong learning and inferring capability [6]. It integrates computer, topology and possibility is a powerful tool for representation and inference of possibility knowledge. It can be used to represent any of the

classic models in IR. Therefore, in this paper, our work is an extension for Markov network information retrieval model.

We first describe the Markov network information model in more details. A Markov network is an undirected graph G and is expressed by G (V, E). Let V be the set of term nodes and E be the set of undirected edges in the graph respectively. The construction of Markov network is similar to previous work [8]. First of all, term correlativity can be measured by mutual information (MI), latent semantic index or term co-occurrences. Considered undirected characteristics of Markov network, our work simply adopts co-occurrences between terms to measure term relationship as follows:

$$S(t_i, t_j) = \frac{N(t_i, t_j)}{N(t_i) + N(t_j) - N(t_i, t_j)} \qquad (1)$$

Where s (t_i, t_j) measures the relationship between t_i and t_j in Markov network. N (t_i, t_j) is the frequency of co-occurrences of t_i and t_j in corpus, and N (t_i), N (t_j) have the definition similar to N (t_i, t_j).

After the relationship between terms is computed in terms of methods above, we construct term space representing the relationships between terms. If the value of relationship s (t_i, t_j) between t_i and t_j is more than a give threshold, then we add an edge between t_i and t_j. It is a learning process to construct term space. Therefore, the Markov network of terms is built as the base graph shown in Fig.1.

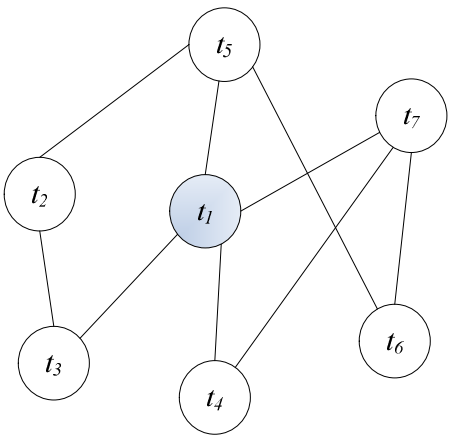

Fig.1 Markov network of terms

III. Query Expansion Based on Quasi-Clique

From the Markov network of terms shown in Fig.1 above, if a term t_1 becomes the center node, its co-occurrence terms become neighboring nodes attached to the center node. According to the previous works that only consider the direct relationship between co-occurrence terms and t_1 such as cliques, candidates of t_1 are gotten as $\{t_3, t_4, t_5, t_7\}$. Note that terms t_2 and t6 both have many common nodes which are all well connected to t_1.This can be used as a clue to unearth the hidden similarity to the core term t_1. If we look at the relationships existing in the whole Markov network of terms instead of individual vertices in Fig.1, then we are able to find that t_2 and t_6 are with high relevance to t_1. Therefore, in this paper, we adopt the notion of Quasi-Clique to capture contextual information mined from the term network for per query term in addition to syntactic similarity. Our work shares many closely related insights in [9] which take useful of Quasi-Clique to improving grouped-entity resolution. A Quasi-Clique is a good graph structure to identify compact sub-graphs. In our Markov network of term space, a Quasi-Clique is a set of terms that are highly interactive with each other. Therefore, a Quasi-Clique in Markov network can strongly indicates the existence of a same topic. Since a Quasi-Clique contains a group of highly relevant terms, it may be more reliable in expressing a similar concept than individual terms. Therefore, Quasi-Cliques can be used to represent a same topic and it is benefit to add such Quasi-Cliques for a query term expansion to reduce topic drift.

A. *Quasi-Clique.*

The definition of Quasi-Clique is shown as: a connected graph G is a Quasi-Clique graph ($0 < \gamma \leq 1$) if every node in the graph has a degree at least $\gamma * (n-1)$, where n is the number of nodes in graph G, the value of γ indicates the compactness of a Quasi-Clique. As shown in [10], Quasi-Cliques have an interesting property that when γ is not too small, a γ-Quasi-Clique is compact because the diameter of the Quasi-Clique is small. The diameter D (G) of G is computed as:

$$D(G) \begin{cases} = 1 & \text{if } 1 \geq \gamma > \frac{n-2}{n-1} \\ \leq 2 & \text{if } \frac{n-2}{n-1} \geq \gamma \geq \frac{1}{2} \\ \leq 3\lfloor \frac{n}{\gamma(n-1)+1} \rfloor - 3 & \text{if } \frac{1}{2} > \gamma \geq \frac{2}{n-1} \text{ and } \\ & n \bmod (\gamma(n-1)+1) = 0 \\ \leq 3\lfloor \frac{n}{\gamma(n-1)+1} \rfloor - 2 & \text{if } \frac{1}{2} > \gamma \geq \frac{2}{n-1} \text{ and } \\ & n \bmod (\gamma(n-1)+1) = 1 \\ \leq 3\lfloor \frac{n}{\gamma(n-1)+1} \rfloor - 1 & \text{if } \frac{1}{2} > \gamma \geq \frac{2}{n-1} \text{ and } \\ & n \bmod (\gamma(n-1)+1) \geq 2 \\ \leq n-1 & \text{if } \gamma = \frac{1}{n-1} \end{cases} \qquad (2)$$

Therefore, for a given query term q_i, we first select relevant terms from the base Markov network of terms that their similarities to q_i are higher than a give threshold. And then for the second selection, we will choose γ-Quasi-Clique which sets q_i as a core node. When the value of γ is set to 1, a 1-Quasi-Clique graph is a complete graph such as clique which is adopted in the model [7]. As shown in the Fig.1 above, if the value of γ is set to 0.5, then the list of Quasi-Cliques containing t_1 is $\{(t_1, t_2, t_3, t_5), (t_1, t_4, t_6, t_7)\}$.Therefore, we can see that the query term t_1 will get more relevant candidates such as terms t_2 and t_6.Similarly, every query term will get its candidate list of Quasi-Cliques for query expansion.

B. *Query expansion.*

Since a Quasi-Clique contains a group of highly relevant terms, it well expresses a similar topic. Therefore, when a Quasi-Clique is adopted in query expansion, it is benefit to reduce false positives and topic drift. The key step of query expansion is how to take use of the expanded terms. In this paper, we adopted Quasi-Cliques as query context for query expansion. The number of the extended terms can be settled by using query evaluation and query evaluation can be optimized after experimental training. Documents are ranked

in descending order of this score. Then the score can be computed as follows:

$$p(d_j|q) \propto \sum_{t_i \in q} (\lambda p(t_i|q) p(t_i|d_j) + (1-\lambda) \sum_{t_k \neq t_i \wedge t_k \in C(t_i)} p(t_k|q) p(t_k|d_j)) \quad (5)$$

Where $C(t_i)$ is the list of Quasi-Cliques that set term t_i as the center node; t_k is a term that is contained in the same Quasi-Clique of t_i and λ is a smooth parameter.

IV. SUMMARY

This paper proposes an information retrieval expansion model based on Quasi-Clique. Quasi-Clique is used to measure how strong inter-relationships between terms. Our approach exploits contextual information mined from the term network for per query term in addition to syntactic similarity. The proposal in our work is benefit to select more relevant terms for query expansion and to improve recall and precision in information retrieval. In the future work, to validate our proposal on retrieval performance, we will perform experiments on standard corpuses of IR such as TREC, ADI, MED, CACM, and CICS.

Acknowledgements

This work was financially supported by the Jiangxi Natural Science Foundation (20122BAB211032，20112BDE50049), Jiangxi College of Humanities and Social Sciences Fund (JD1164) and Education Reform Project of Jiangxi provincial universities (JXJG-13-10-13).

References

[1] K. Tamsin Maxwell and W. B. Croft, “Compact Query Term Selection Using Topically Related Text,” In Proc. of SIGIR 2013, pp. 583-592, 2013.

[2] Jianfeng Gao, Gu Xu and Jinxi Xu, “Using Query Expansion Using Path- Constrained Random Walks,” In Proc. of SIGIR 2013, pp. 563-572, 2013.

[3] Y. Li, W. P. R. Luk, K. S. E. Ho and F. L. K. Chung, “Improving weak ad-hoc queries using Wikipedia as external corpus,”In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007. pp. 797-798,2007.

[4] Gelfand B, Wulfekuler M and Punch W F, “Automated concept extraction from plain text,” AAAI 1998 Workshop on Text Categorization. pp. 13-17,1998.

[5] Donald Metzler and W. Bruce Croft, “Latent Concept Expansion Using Markov Random Fields,”SIGIR’07, Amsterdam, The Netherlands, ACM, 2007.

[6] Gan Lixin,Tu Wei, Liu Guodong and Yu Junying, “Review and Perspective of Markov Network Information Retrieval Model,”In Proc of WICOM 2011,Shanghai,China,2011.

[7] Lixin Gan, Shengqian Wang, Mingwen Wang , Zhihua Xie, Lin Zhang and Zhenghua Shu, “Query Expansion based on Concept Clique for Markov Network Information Retrieval Model,”In Proc of the 5th International Conference on Fuzzy Systems and Knowledge Discovery,2008.

[8] Gan Lixin, “The Information Retrieval Model based on Markov concept ,”Nan chang: Jiang Xi Normal University, 2007.

[9] On, B. W., Elmacioglu, E., Lee, D., Kang, J. and Pei, J. , “ Improving grouped-entity resolution using quasi-cliques,”In Data Mining,2006. ICDM'06. Sixth International Conference on.IEEE,pp. 1008-1015,2006.

[10] J. Pei, D. Jiang, and A. Zhang, “On Mining Cross-Graph Quasi-Cliques,” In ACM KDD, 2005.