Grey Relational Analysis for Query Expansion

Junjie Zou, Zhengtao Yu*, Huanyun Zong, Jianyi Guo, and Lei Su

School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, 650051, China
Intelligent Information Processing Key Laboratory,
Kunming University of Science and Technology, Kunming, 650051, China
ztyu@hotmail.com

Abstract. For one-sidedness of the various qualitative expansion methods, we propose a query terms selection method based on Grey Relational Analysis (GRA).We called the fusion expansion technique with GRA (FET-GRA). It calculates weight of expansion term by varied qualitative expansions and comprehensive weight by FET-GRA and thus extracts expansion term in terms of the weight. The experiment result of TREC dataset shows the method (FET-GRA) is substantially superior to TF-IDF, Mutual Information , Local Context Analysis.

Keywords: query expansion, grey relational analysis, FET-GRA.

1 Introduction

Query expansion is the critical process to improve precision of retrieval. At present, there are many effective qualitative expansion methods[1,2,3], however, they all base on a single theory to calculate and select candidate expansion terms, such as TF-IDF, Mutual Information (MI), Local Context Analysis (LCA) which have certain one-sidedness. For example, TFIDF on the assumption of independent term is good to simple query regardless semantic problem but poor performance to query expansion which requires context, on the contrary of LCA. Grey Relational Analysis (GRA) is a part of Grey System Theory[4]. It can offer a new evaluation of estimate after integrating the relation of each qualitative method. Taking expansion of airlines that currently use Boeing 747 planes as example, candidate term Singapore gets low and common scores in TFIDF and MI but a good score in LCA. GRA can be used in the decision problem whether Singapore is a good expansion term. It integrates and calculates the evaluations of estimate according to the three differences and set-term topology and other information. The evaluations estimate the importance of Singapore. The thesis uses Relational Analysis method[5] to integrate and revalue query terms weights from different qualitative query expansion methods. It thus extracts and expands the optimal expansion term.

This paper is organized as follows: Section 2 describes related work. Section 3 presents the query expansion and grey relation theory; Section 4 combines the

* Corresponding author.

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 386–393, 2013.

Grey Relational Analysis with query expansion. In Section 5, we describe our experimental methodology and the results as well as analysis.

2 Related Work

Early scholars proposed the query expansion technology based on global analysis, it holds that there were correlations between words in a corpus, which reflected by the co-occurrence times they appeared in the corpus. Latent semantic analysis (LSA)[6] is an earlier proposed global method; its core idea is to map high-dimensional vector space into low-dimensional latent semantic space by singular value decomposition. It may be more appropriate for system with small corpus, but for larger or even web data, the feasibility of LSA method comes into a serious challenge, because the web data is too large to make global analysis. Previous work[7,8] is also a kind of global analysis technology based on document, which selects the words with higher word frequencies by statistics as expansion terms added to the original query. This method also has a great weakness in big computation cost.

Local analysis technology usually consists of two steps. First step, use the original query to obtain the original search results from retrieval system, and select the top N documents of original initial search results as local documents collection D. Second step, take out the top ranked candidate words analyzed in D, which added to the original query to reconstruct the query. Atter and Fraenkel first proposed the idea of local analysis in the literature[9], Xu and Croft in the literature[1] further refined local analysis technology and put forth the local context analysis method , rank concepts according to calculate the terms similarity between original query and local documents collection, the top ranked concepts are added to original query to expand query. On this basis, Sun et al in the literature[10,2] used Google for initial retrieve, later expanded query with local context analysis method and obtained certain effects. However, these methods also have the corresponding problems, for example, the top N documents retrieved by initial query is not quite constant with user requirements, whereas the query terms expanded from those methods far from improve retrieval accuracy ,it would weaken retrieval performance.

In terms of Language model, since Ponte and Croft first introduced it into the field of information retrieval[11], it has been widely used. Bai et al in the literature[12] made a further research on Language model, focused on how to take advantage of hyper-spatial analysis of Linguistics and co-occurrence frequency methods to calculate the probability of words to query model, used the high probability of words to expand query. In addition, Collins-Thompson et al used random walk model[13] to expand the query expansion, Cui et al conducted a personalized query expansion based on user log[14].

The above methods each possess its advantages and disadvantages, how to improve the accuracy of the query expansion with these different methods is the key to the study. This paper presented a fusion evaluation technique integrated the existing expansion methods by using grey relational analysis.

3 Query Expansion and Grey System Theory

In order to verify the effectiveness of the fusion expansion technique with GRA (FET-GRA), this paper chosen three basic expansion methods for fusion evaluation. What follows is a brief sketch of these three basic expansion methods, and describes the core theory of fusion expansion technique with GRA (FET-GRA).

3.1 Query Expansion Methods

The idea in Local context analysis[1] is that noun groups are used as concepts and concepts are selected based on co-occurrence with query terms. Calculate and rank the concepts according to the relevance between concepts and query, the top ranked concepts are chosen as expansion terms. The concepts of context are similar to local feedback, relevance calculation used the original top ranked N documents in traditional feedback technology, but the best passage are used instead of whole document in Local context analysis technology. Local context analysis technology is a practical technology, which combines global analysis and local feedback, often used in query expansion.

We adopt a similar method to literature [1,2] for expanding query Q in tourism domain. The first thing is to determine context of the paragraphs set SP, use Google to retrieve the top N information fragments collection $SP = \{s_i, i = 1, ..., n\}$ segment sentence for each s_i.Then, calculate the relevance $SIM(Q, c)$ between each concept and query Q using paragraphs collection SP, the calculation formula is as follows:

$$SIM(Q,c) = \frac{1}{Z} \cdot \prod_{t_i \in Q} \left\{ \delta + \frac{\log\left[\sum_{j=1}^{n} (tf_{ij} \cdot tf_{cj})\right] \times \log(\frac{N}{N_c})}{\log(n)} \right\}^{\log(\frac{N}{N_i})} \tag{1}$$

Where Z is normalized factor, denotes smoothing factor for preventing the equation is zero, tf_{ij} and tfcj stand for word frequency of t_i and concept c in paragraph SP, N is the total number of paragraphs in paragraph collection, N_i and Nc are the number of t_i and concept c appeared in paragraph collection.

Next, rank the above results of calculation. Finally select the top-k concepts as candidate words added to the original query. For meaningful on words ranking, we use Indri query language of Indri retrieval platform to refectory the query, the refactor query expression is as #weight($w_0 q_1 ... w_0 q_m w_1 c_1 ... w_k c_k$).Where q_i indicates the key words of the original query Q, c_i indicates the i^{th} ranked concept, w_i denotes the weight of key words in refactor query.w_i calculation method is shown in Equation (2).

$$w_i = \begin{cases} 2.0, i = 0 \\ (k - 0.9 \times i)/k, \text{else} \end{cases} \tag{2}$$

For the query expansion method based on TF-IDF, its core TF-IDF[15].Based on the original mutual information[16], we use an improved method to calculate the

mutual information of candidate words w_i and query Q, the details are shown in Equation (5).Where m is the number of keywords, Z_m is normalized factor,δ is anti-zero factor, $\delta = 0.01$ in this paper.

$$I(w_i : Q) = \frac{1}{Z_m} \cdot \sum_{t=1}^{m} \left(\left[\log_2 \frac{P(w_i, q_t)}{P(w_i) \cdot P(q_t)} \right] + \delta \right) \tag{3}$$

3.2 Grey Relation Analyses

Grey relation analysis (GRA) was pioneered by Deng Julong in 1984.It used to solve these problems, having incomplete running mechanism, lacking of behavior data, devoid of experience in treatment, being naked to inherent connotation. We introduced relational definition about GRA in this section.

For the elements(factors) between two systems, the measurement of relevance changed over time or different objects called Grey Relational Grade. In the course of system development, if two elements have much consistency in developmental trends, namely high degree of synchronous changes, which can be described as a high-related degree of two factors, on the contrary, it is relatively low. Consequently, grey relational analysis method established on the similarity or diversity of developmental trends of these elements, namely Grey Relational Grade, as a measurement approach of related degree of these elements. Grey system theory proposed the concept of grey relational analysis for each subsystem with an intension to seek the numerical relations between every subsystems (or elements) by certain means. Therefore, gray relational analysis is a kind of quantitative description and comparison for the developmental trends of a system, its basic idea is to determine the similarity degree of geometric figures between reference sequence and several compare sequences to judge whether closely related, which reflected the correlation between curves.

This paper regarding different expansion methods as different systems, each system can score a certain candidate term, and takes data sequence formed by query terms as reference sequence, other data sequence formed by candidate terms as compare sequence. Hence, we can calculate the similarity of reference sequence and compare sequence by using gray relational analysis.

4 Fusion Expansion Technique with GRA

The modeling procedure of Fusion expansion technique GRA(FET-GRA) method. Firstly, we use several qualitative query expansion methods to generate feature matrix of candidate terms compare sequence. Secondly, GRA method requires the optimal collocation of feature value, so it requires to confirm a feature reference sequence. Thirdly, to calculate the grey relational coefficient between feature reference sequence and candidate terms compare sequence, thereafter solve the relational sequence of candidate terms and rank them. Lastly, we can then extract the candidate terms ranking higher as an expansion terms.

The detailed modeling process of GRA method in query expansion is as follows.

Step 1: Build model grade feature matrix $\Gamma(\mathcal{M})$. We use longitudinal vector $\mathcal{M} = \{m1, m2, m3\}$ of feature matrix from TFIDF, LCA and improved MI with weight factors, and choose the candidate terms from relative documents as horizontal vector to construct Equation (4).

$$\Gamma(\mathcal{M}) = [m_i(k)]_{n\times d}, \qquad where\ i = 1\ to\ n; k = 1\ to\ d \tag{4}$$

Where $m_i(k)$ denotes the value of the i^{th} candidate term from the k^{th} qualitative method, $d = 3$ in our paper.

Step 2: Construct reference sequence. We design an auto-selected optimal reference sequence method. We assume the query input by users has a positive effect on the result while does not digress expansion topic. Suppose $\mathcal{Q} = \{q_1, q_2, \ldots, q_w\}$ be a partition of query keywords, the method for constructing reference sequence $\mathcal{V}$ by using query $\mathcal{Q}$ as shown in the formula(5).

$$\mathcal{V} = [\max_k \{m_i(k)\} \mid i \in \mathcal{Q}]_{1\times d}, \qquad where\ k = 1\ to\ d \tag{5}$$

Step 3: Generate n×d relational matrix $\mathcal{Z}$. Firstly, calculate the grey relational coefficient (GRC) $\zeta_i(k)$ of each candidate term corresponding to comparison sequence and reference sequence respectively. Secondly, compose matrix $\mathcal{Z}$ using $\zeta_i(k)$, where the meaning of horizontal vector and longitudinal vector is similar to step 1 . The details are shown in Equation (6).

$$\begin{aligned}
&\mathcal{Z} = [\zeta_i(k)]_{n\times d}, \zeta_i(k) = \frac{\min_i \min_k \mathcal{CS}_{ik} + \rho \times \max_i \max_k \mathcal{CS}_{ik}}{\mathcal{CS}_{ik} + \rho \times \max_i \max_k CS_{ik}} \\
&\zeta_i(k) \ \text{ is a GRC based with } \ \mathrm{q_i} \text{ and } \mathcal{V}(\mathrm{k}). \\
&\mathcal{CS}_{ik} \ \text{ is a absolute values of } \mathcal{V}(\mathrm{k}) - \mathrm{m_i(k)}. \\
&\mathcal{V}(k) \ \text{ is kth feature value of reference vector } \mathcal{V}. \\
&\quad i \ \text{ is a term suffix that it is a value of 1 to n.} \\
&\quad \rho \ \text{ is a distinguishing coefficient.} \\
&\quad k \ \text{ is a feature suffix that it is a value of 1 to d.}
\end{aligned} \tag{6}$$

Step 4: Calculate the grey relational sequence γ_i of candidate terms. First, calculate γ_i using grey relational coefficient.Then, rank the candidate terms according to γ_i, and select the first n with maximal value γ_i of candidate terms as expansion terms. Equation (7) illustrates the calculation of γ_i.

$$\gamma_i = \frac{1}{d}\sum_{k=1}^{d} \mathcal{W}_k \cdot \zeta_i(k),\ \mathcal{W}_k\ is\ weight\ of\ feature \tag{7}$$

5 Experiments

We conducted experiments to verify the effectiveness of query expansion using FET-GRA.In this section,we first introduce the data sets used in experiments. Then we demonstrate the effectiveness of our approach(FET-GRA) in query expansion.

5.1 Data Set

The purpose of the experiments is to evaluate the performance of query expansion by FET-GRA method. We conduct the experiments in TREC ClueWeb09 (category B set), and build the retrieval experiment platform with the help of Lemur toolkit (www.lemurproject.org). We construct a query set including 50 questions. It consists of two parts. Part 1 is TREC2009 Entity Track's query set. Only the entity name are used as queries. Part 2 is the queries that were the query log of a commercial web search engine. The top-100 snippets are extracted by Google retrieval and map with the retrieval result of Lemur toolkit.

5.2 FET-GRA Experiments

In our experiments,TF-IDF,LCA and MI with weight's factor make up the fusion expansion set.Then we use fusion expansion technique with GRA(FET-GRA) and distinguishing coefficient $\rho = 0.5$ of GRA query expansion method, relational sequence $\mathcal{W}_k = 1$,namely, without weighting. Table 1 illustrates the results of the experiments. The value outside of bracket indicates relevant precision and the one inside is the improved rate compared to the baseline methods.

Table 1. Precision of different expansion methods

%	TF-IDF	MI	LCA	FET-GRA
p@5	55.6	70.0(+25.9)	75.4(+35.6)	92.2(+62.2)
p@10	53.4	65.5(+22.7)	57.5(+7.7)	75.6(+41.6)
p@20	43.8	45.5(+3.9)	51.3(+17.1)	66.4(+51.6)
p@30	37.5	47.5(+26.7)	49.6(+32.3)	60.4(+61.1)
p@40	31.8	32.6(+2.5)	42.5(+33.6)	52.8(+66.0)
p@50	26.0	28.5(+9.6)	36.3(+39.6)	43.3(+66.5)
p@60	22.9	23.7(+3.4)	31.1(+35.8)	38.3(+67.2)
average	38.7	44.8(+26.1)	49.1(+26.9)	61.0(+57.6)

From the experimental results in Table 1, we can see the precision of MI, LCA and FET-GRA improves over TF-IDF method. The performance of MI and LCA improve 26.1% and 26.9% respectively while the one of FET-GRA improves obviously 57.6%. In order to validate the relevance of different expansion methods, we use Spearman's rank correlation test [17], take the precision of Table 1 as a test sample, four methods as random variables , and then test FET-GRA with other three methods respectively, In consequence, we obtain the following values: $P - value_{TFIDF} = 3.97 \times 10^{-4}$, $P - value_{MI} = 6.74 \times 10^{-3}$, $P - value_{LCA} = 3.97 \times 10^{-4}$. Each $P - value$ is less than $\alpha = 0.01$, so it indicates that the random variables have correlation, namely, FET-GRA method improves expansion capability under the premise of keeping to qualitative methods.

Figure 1 illustrates the analytical result of stable improved rate of MI, LCA and FET-GRA by using the improved rate data in Table 1.

As can be seen from Figure 1, FET-GRA has higher robustness as well as a better improved rate compared to the baseline methods than LCA's and MI's.

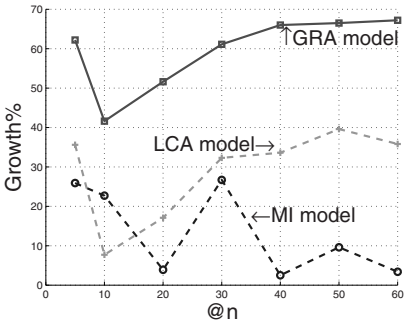

Fig. 1. Growth Curve of three method

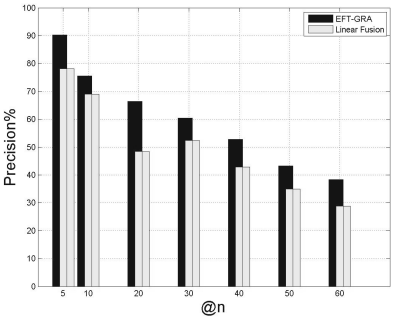

Fig. 2. FET-GRA v.s. Linear

6 Conclusions

We apply FET-GRA method to query expansion. It can effectively select the query relational terms and also improve the precision of query expansion as well as robustness. In the future work, we will study the inaccurate reference sequence caused by query drift problem.

Acknowledgments. This paper is supported by National Nature Science Foundation (No.61175068), and the Open Fund of Software Engineering Key Laboratory of Yunnan Province (No.2011SE14), and the Ministry of Education of Returned Overseas Students to Start Research and Fund Projects.

References

1. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: 19th ACM SIGIR, pp. 4–11. ACM Press, Zurich (1996)
2. Sun, R., Ong, C.H., Chua, T.S.: Mining dependency relations for query expansion in passage retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 382–389. ACM Press, Washington (2006)
3. Nen-Townsend, S., Zhou, Y., Croft, W.B.: A framework for selective query expansion. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 236–237. ACM Press (2004)
4. Deng, J.L.: Control problems of grey systems. Systems & Control Letters 1(5), 288–294 (1989)
5. Deng, J.L.: Introduction to grey system theory. The Journal of Grey System 1(1), 1–24 (1989)
6. Deerwester, S.C., Dumais, S.T., Landauer, T.K., et al.: Indexing by latent semantic analysis. JASIS 41, 391–407 (1990)
7. Jing, Y., Croft, W.B.: An association thesaurus for information retrieval. In: Proceedings of RIAO, pp. 146–160 (1994)

8. Callan, J.P., Croft, W.B., Broglio, J.: TREC and TIPSTER experiments with INQUERY. Information Processing & Management 31(3), 327–343 (1995)
9. Attar, R., Fraenkel, A.S.: Local feedback in full-text retrieval systems. Journal of the ACM (JACM) 24(3), 397–417 (1997)
10. Cui, H., Sun, R., Li, K., et al.: Question answering passage retrieval using dependency relations. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 382–389. ACM Press, Salvador (2005)
11. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281. ACM Press, Melbourne (1998)
12. Bai, J., Song, D., Bruza, P., et al.: Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 275–281. ACM Press, Bremen (2005)
13. Collins-Thompson, K., Callan, J.: Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 704–711. ACM Press, Bremen (2005)
14. Cui, H., Wen, J.R., Nie, J.Y., et al.: Probabilistic query expansion using query logs. In: Proceedings of the 11th International Conference on World Wide Web, pp. 325–332. ACM Press, Honolulu (2002)
15. Liu, Y., Ciliax, B.J., Borges, K., et al.: Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering. In: Computational Systems Bioinformatics Conference 2004, pp. 394–404. ACM Press, Honolulu (2004)
16. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
17. Gibbons, J.D., Chakraborti, S.: Nonparametric statistical inference. CRC press (2003)