Choosing meaningful structure data for improving web search

Guo Xi Yang Xiaochun Yu Ge Li Guangao

(College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

Abstract: In order to improve the quality of web search, a new query expansion method by choosing meaningful structure data from a domain database is proposed. It categories attributes into three different classes, named as concept attribute, context attribute and meaningless attribute, according to their semantic features which are document frequency features and distinguishing capability features. It also defines the semantic relevance between two attributes when they have correlations in the database. Then it proposes trie-bitmap structure and pair pointer tables to implement efficient algorithms for discovering attribute semantic feature and detecting their semantic relevances. By using semantic attributes and their semantic relevances, expansion words can be generated and embedded into a vector space model with interpolation parameters. The experiments use an IMDB movie database and real texts collections to evaluate the proposed method by comparing its performance with a classical vector space model. The results show that the proposed method can improve text search efficiently and also improve both semantic features and semantic relevances with good separation capabilities.

Key words: web; semantic; attributes relationship; structure data; query expansion

A search engine finds relevant texts containing keywords provided by users based on information search models[1]. However, these keywords are often insufficient and imprecise[2]. This problem results in irrelevant texts being returned and relevant texts being lost. Query expansion[3] improves the descriptive capability of keywords by adding semantically relevant words to original keywords implicitly[4] or explicitly[5]. Expansion words are generated by analyzing their semantic relationships with original keywords. Traditionally, semantic relationships are stored in three kinds of sources, such as thesaurus[6], co-occurrence[7] and query logs[8]. Currently, there are some new variations of query expansion methods[9–10]. Moreover, some literature focuses on discovering relationships between structure data and text[11–12].

In this paper, we construct a new source of words' semantic relationships based on a domain database which has not been utilized as a semantic relationship provider. We use attribute values and their semantic relationships in structure data to generate expansion words by defining attributes' semantic features and analyzing semantic relevances between two attributes. And we propose efficient algorithms to discover attributes' semantic features and detect their semantic relevancies between two attributes by sampling and estimating. Then we change the vector space information retrieval model to embed expansion words.

1 Attribute Semantic Feature and Semantic Relevance

We class attributes into three different categories (concept attribute, context attribute and meaningless attribute) by observing the features of sampled attribute values. And we detect and evaluate semantic relationships between two attributes.

1.1 Attribute semantic feature

We can class an attribute by analyzing its document frequency features and by distinguishing capabilities as shown in Fig. 1. A concept attribute with low document frequency and high distinguishing capability contains values representing domain entities. A context attribute with high document frequency contains values describing domain entities. A meaningless attribute with low document frequency and low distinguishing capability contains values that are not relevant to the specific domain. For example, in Fig. 2, values of Moviename represent entities in the movie domain, so Moviename is a concept attribute. Values of Genre and Playdate describe movie entities, so Genre and Playdate are context attributes. Values of Plotby (who adds this movie record) have no semantic meaning in the movie domain, so Plotby is a meaningless attribute.

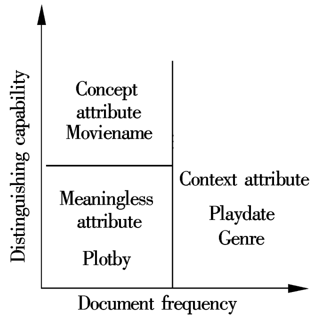
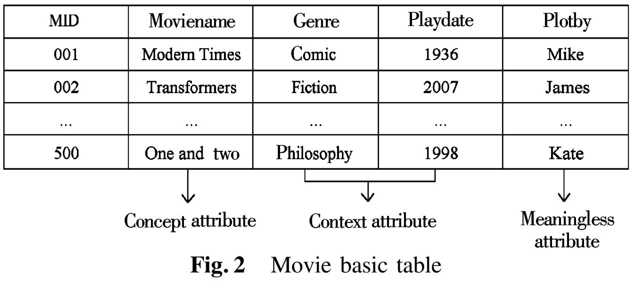


Fig. 1 Three attribute categories



MID	Moviename	Genre	Playdate	Plotby
001	Modern Times	Comic	1936	Mike
002	Transformers	Fiction	2007	James
...	...	...	...	...
500	One and two	Philosophy	1998	Kate

Concept attribute
Context attribute
Meaningless attribute

Fig. 2 Movie basic table

Received 2008-04-15.

Biographies: Guo Xi (1983 —), female, graduate; Yang Xiaochun (corresponding author), female, doctor, associate professor, yangxc@mail.neu.edu.cn.

Foundation items: Program for New Century Excellent Talents in University (No. NCET-06-0290), the National Natural Science Foundation of China (No. 60503036), the Fok Ying Tong Education Foundation Award (No. 104027).

Citation: Guo Xi, Yang Xiaochun, Yu Ge, et al. Choosing meaningful structure data for improving web search[J]. Journal of Southeast University (English Edition), 2008, 24(3): 343 – 346.

Document frequency[11] f_i of an attribute value possibly frequent interval ($0.01 \leqslant f_i \leqslant 0.1$) and stop word[11] interval ($f_i > 0.1$). The predominant frequency interval (PFI) covers most f_i of the attribute values. An attribute's document frequency feature F can be described by the expectation of f_i in PFI:

$$E(F) = \sum_{i=1}^{t} f_i p_i \qquad p_i = \Pr\{F = f_i\}; \quad i = 1, 2, \ldots, t \tag{1}$$

where t is the number of f_i in PFI. The estimation of F is

$$\hat{F} = \frac{1}{t'} \sum_{i=1}^{t'} O_i = \overline{O} \tag{2}$$

where o_i is the observation of f_i in the sample's PFI and t' is the number of o_i.

If an attribute has a high distinguishing capability, appearances and absences of its values are highly dependent on the text collection's specific domain. We use a 2 × 2 contingency table to test its dependence and to estimate its distinguishing capability. The difference between our method and feature selection[11] is that our method is based on observation of attribute values, which is a set of infrequent words. We evaluate the dependence of attribute A and the specific domain D by test statistics:

$$T(D, A) = \sum_{i \in (0,1)} \sum_{j \in (0,1)} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

where O_{ij} represents the number of texts containing words in attribute A's value set and E_{ij} is $(N_i \times M_j)/N$ in which N is the total number of texts in the domain and other collections. The distribution of $T(D, A)$ can be estimated by χ^2 distribution with a freedom of 1. If $T(D, A)$ is more than $1 - \alpha$ quantile of χ^2 distribution, we can reject the hypothesis that they are independent. The greater $T(D, A)$ is, the higher an attribute's distinguishing capability is.

1.2 Attribute semantic relevance

Two attributes have correlations when they are in the same table or they are in different tables but joined by foreign keys. By analyzing the semantic relevances of the attributes' value pairs we can detect the semantic relevances between two attributes. A value pair set of attributes A and B is $P = \{\langle a, b \rangle \mid A \infty B, a \in A, b \in B\}$, where $A \infty B$ represents that they have relationships in the database and that a, b are their values. The semantic relevance r of value pair $\langle a, b \rangle$ is $f_{ab}/(f_a \times f_b)$ where f_{ab} represents the number of texts containing both a and b. Attribute semantic relevance R is the expectation of r and can be estimated by observed semantic relevances of sampled value pairs.

1.3 Feature discovery and relevance detection

We propose a Trie-Bitmap and a pair pointer table to discover attributes' semantic features and detect their semantic relevances. We store all the distinct values of the attributes into trie and mark every value's terminative node by bitmap in which every bit corresponding to its appearance is 1 and its absence is 0 in a given text. In order to find values' appearances in the text collection we use the AhoCorasick algorithm[13] to solve such dictionary exact match problems efficiently. Then f_i of every sampled value is the number of 1s in its bitmap. We use the pair pointer table to store relationships and pointers of value pairs from two attributes. Co-occurrences of value pairs can be computed by intersecting their corresponding bitmaps.

2 Improve Text Search by Using Semantic Attribute Values

We parse original query keywords to obtain their semantic meaning and map them into concept or context attributes in order to obtain their expansion words, then we embed expansion words into the vector space model.

2.1 Parse and map query keywords

A query is decomposed into a terms set in order to cover the semantic meaning of the query. Terms are generated by sliding a variable length (between one and the number of keywords) window on the keywords' sequence. Then we match these terms into concept attributes or context attributes. The optimal term combination of a query is that the distance between the two far-most semantic relevant terms is minimal. If we cannot find an optimal term combination, user-assisted query expansion[14] will be carried out.

2.2 Embed expansion words into vector space model

Expansion words generated are embedded into the classical vector space model with interpolation parameters:

$$\boldsymbol{q}' = \alpha \boldsymbol{q} + \beta \sum_{i=1 \ldots |c|} \boldsymbol{e}_i \tag{4}$$

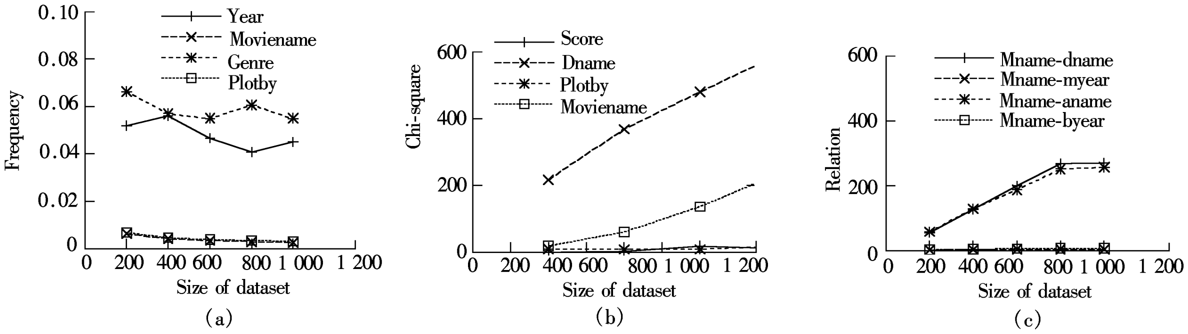
where $\boldsymbol{e}_i$ represents semantically relevant words of the term from the optimal term combination c; α and β are two interpolation constants used to adjust the influences of original query keywords and expansion words.

3 Evaluation

In the experiments, we use IMDB (700 MB with 20 tables) as the domain database, 1 000 movie texts from the website New York Times and the Greatest Films, and 1 000 other texts from 20 Newsgroups. Algorithms are implemented by C++ and run on a PC(Intel Pentium R4 CPU 2.40 GHz, 512 MB memory).

Fig. 3(a) shows the document frequency feature of four attributes which can be separated into context attributes and concept or meaningless attributes. Fig. 3(b) shows the distinguishing capability of attributes with low document frequency which can be separated into concept and meaningless attributes. Fig. 3(c) shows semantic relevances between Moviename and four other attributes. And there are two attributes relevant to it. Moreover, the separation capabilities of the document frequency feature, distinguishing capability feature and semantic relevance increase when accompanied by an increase in the dataset. In Fig. 3, "dname" is the director name, "aname" is the actor name, "mname" is the movie name, "myear" is the year the movie is put on and "byear" is the year the director was born.

We evaluate the proposed query expansion method (PNM + EP) by comparing its precision, recall, accuracy, and average score with a pivoted normalization model (PNM)[1]. The query dataset is made up of insufficient and imprecise queries. Tab. 1 shows that by using PNM + EP, precision and recall are improved a lot with a little decrease in accuracy. An increase in the average score will result in better ranking of relevant texts.



(a) (b) (c)

Fig. 3 Evaluation results. (a) Document frequency; (b) Distinguishing capability; (c) Semantic relevance

Tab. 1 Performance comparison

Precision/%		Recall/%		Accuracy/%		Average score	
PNM	PNM + EP	PNM	PNM + EP	PNM	PNM + EP	PNM	PNM + EP
9.97	15.4	40	80	96.13	95.97	2.95	8.34

4 Conclusion

In this paper, we propose a new query expansion method by using structure data in a domain database to improve text search. And we define attribute semantic features including document frequency features and distinguishing capabilities. We also define semantic relevances between two attributes. Then we give efficient algorithms to discover semantic features and detect semantic relevances. We embed generated expansion words into a vector space model and evaluate the performance of the proposed method. However, generated expansion words can also be embedded into other information retrieval models and we leave that for the future work.

References

[1] Manning Christopher D, Raghavan Prabhakar, Schutze Hinrich. *An introduction to information retrieval* [M]. Cambridge: Cambridge University Press, 2008: 109 – 133; 253 – 287.

[2] Billerbeck Bodo, Zobel Justin. Questioning query expansion: an examination of behavior and parameters [C]//*Proc of the Fifteenth Australasian Database Conference*. Dunedin, New Zealand, 2004: 69 – 76.

[3] Custis Tonya, Al-Kofahi Khalid. A new approach for evaluating query expansion: query-document term mismatch [C]// *Proc of the* 30*th Annual International ACM SIGIR Conference*. New York: ACM Press, 2007: 575 – 582.

[4] Cao Guihong, Nie Jianyun, Bai Jing. Integrating word relationships into language models [C]//*Proc of the* 28*th Annual International ACM SIGIR Conference*. New York: ACM Press, 2005: 298 – 305.

[5] Crouch Carolyn J, Yang Bokyung. Experiments in automatic statistical thesaurus construction [C]//*Proc of the* 15*th Annual International ACM SIGIR Conference*. New York: ACM Press, 1992: 77 – 88.

[6] Park Laurence A F, Ramamohanarao Kotagiri. Query expansion using a collection dependent probabilistic latent semantic thesaurus [C]//*Proc of the* 11*th Pacific-Asia Conference, PAKDD*. Nanjing, China, 2007: 224 – 235.

[7] Fang Hui, Zhai Chengxiang. Semantic term matching in axiomatic approaches to information retrieval [C]//*Proc of the* 29*th Annual International ACM SIGIR Conference*. New York: ACM Press, 2006: 115 – 122.

[8] Fonseca Bruno M, Golgher Paulo, Possas Bruno, et al. Concept-based interactive query expansion [C]//*Proc of the* 14*th ACM International Conference on Information and Knowledge Management*. New York: ACM Press, 2006: 696 – 703.

[9] Nandi Arnab, Jagadish H V. Effective phrase prediction [C]//*Proc of the* 33*rd International Conference on Very Large Data Bases*. Vienna, Austria, 2007: 219 – 230.

[10] Bast Holger, Weber Ingmar. Type less, find more: fast autocompletion search with a succint index [C]//*Proc of the* 29*th Annual International ACM SIGIR Conference*. New York: ACM Press, 2006: 364 – 371.

[11] Chakaravarthy Venkatesan T, Gupta Himanshu, Roy Prasan, et al. Efficiently linking text documents with relevant structured information [C]//*Proc of the* 32*nd International Conference on Very Large Data Bases*. Seoul, Korea, 2006: 667 – 678.

[12] Chakrabarti Kaushik, Ganti Venkatesh, Han Jiawei, et al. Ranking objects based on relationships [C]//*Proc of the* 2006 *ACM SIGMOD International Conference on Management of Data*. New York: ACM Press, 2006: 371 – 382.

[13] Gusfield Dan. *Algorithms on strings, trees, and sequences: computer science and computational biology* [M]. New York: Cambridge University Press, 1997: 16 – 67.

[14] Bodner Richard C, Song Fei. Knowledge-based approaches to query expansion in information retrieval [C]//*Proc of the* 11*th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*. Springer-Verlag, 1996: 146 – 158.

用于改善 web 搜索的结构化数据抽取技术

郭 茜 杨晓春 于 戈 李广翔

（东北大学信息科学与工程学院，沈阳 110004）

摘要：为了提高 web 文本搜索质量，提出了基于语义结构化数据的查询扩展方法. 通过分析属性的语义特征（文档频率特征和辨识能力特征）将属性分为概念属性、背景属性和无用属性 3 类，并且提出了衡量属性语义相关度的标准. 设计了 trie-bitmap 和 pair pointer table 数据结构来实现发掘属性语义特征和检测属性语义相关度的有效算法. 通过使用合适的属性和它们的语义关系，可以为查询关键字生成扩展词并将它们嵌入到具有插值参数的向量空间模型中. 实验使用 IMDB 电影数据库和真实文本数据集来比较所提方法和原始向量空间模型的性能. 实验结果证明所提出的查询扩展方法可以有效地提高文本搜索性能，同时属性语义特征和属性语义相关度都具有良好的分类能力.

关键词：web；语义；属性关系；结构化数据；查询扩展

中图分类号：TP311