Journal of Southeast University (English Edition) Vol. 24, No. 3, pp. 312 – 314 Sept. 2008 ISSN 1003—7985

Latent semantic analysis for query interfaces of deep web sites

Mao Qinjiao[1] Feng Boqin[1] Pan Shanliang[2]

([1]Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China)

([2]Information Science and Engineering Institute, Ningbo University, Ningbo 315211, China)

Abstract: To further enhance the efficiencies of search engines, achieving capabilities of searching, indexing and locating the information in the deep web, latent semantic analysis is a simple and effective way. Through the latent semantic analysis of the attributes in the query interfaces and the unique entrances of the deep web sites, the hidden semantic structure information can be retrieved and dimension reduction can be achieved to a certain extent. Using this semantic structure information, the contents in the site can be inferred and the similarity measures among sites in deep web can be revised. Experimental results show that latent semantic analysis revises and improves the semantic understanding of the query form in the deep web, which overcomes the shortcomings of the keyword-based methods. This approach can be used to effectively search the most similar site for any given site and to obtain a site list which conforms to the restrictions one specifies.

Key words: deep web; information retrieval; latent semantic analysis; singular value decomposition

The world wide web can be divided into the surface web and the deep web. The deep web, which is also referred to as the invisible or hidden web, is related to the surface web. It can be simply considered as information stored in the database, and only through the corresponding dynamics response to the request in a form can these data be retrieved. Bergman[1] pointed out that the information in the deep web is approximately 500 times greater than that in the surface web. Today, only 0.03% of the content on the web can be searched. To further enhance the efficiency of search engines, we need to achieve capabilities of searching, indexing and locating the information in the deep web.

The deep web sites contain heterogeneous database sources with query interfaces as their unique entrances. Some what noteworthy, the query interfaces provide us the greatest information[2]. Through a form, we are able to identify the corresponding database data. We analyze form pages based on some of the current retrieval technologies. Traditional retrieval technologies based on keyword queries have achieved good results on the surface web but they are not suitable here since developers may use different words to describe the same thing. Human beings can distinguish the meaning of the form by the meanings of the words while machines cannot. In this paper, we apply latent semantic analysis (LSA) to the deep web to locate data sources, aiming to obtain the potential semantic relationships of forms and attributes in the forms. The method can be used to effectively search the most similar site for any given site; or to obtain a site list which conforms to the restrictions one specifies.

Received 2008-04-15.

Biographies: Mao Qinjiao (1983—), female, graduate; Feng Boqin (corresponding author), male, professor, bqfeng@mail.xjtu.edu.cn.

Citation: Mao Qinjiao, Feng Boqin, Pan Shanliang. Latent semantic analysis for query interfaces of deep web sites[J]. Journal of Southeast University (English Edition), 2008, 24(3): 312 – 314.

1 Related Work

Research on the deep web has been going on for 10 years, mostly dedicated to research and solutions in pattern matching, query rewriting, and so on. Zhang et al.[3] focused on query conversion, and developed a lightweight field-based form assistant. Kabra et al.[4] developed a system which can find appropriate sources for users by allowing them to input just an imprecise initial query. Some studies have tried to make classifications of the data resources based on the analyses of the results returned by submitting some random queries[5–6]. Raghavan et al.[7] proposed a hidden web crawling framework which mainly studied the hidden web interface rather than automatic query probes. On the contrary, Ntoulas et al.[8] studied generating automatic queries without any interaction. LSA is also a popular study point in information retrieval[9–10], mainly to address some of the issues caused by semantics.

2 Theoretical Basis for Latent Semantic Analysis

LSA can be used to automatically carry out knowledge acquisition and expression[11]. It makes a statistical analysis on a large number of text sets to extract the contexts and the meanings of words, and, therefore, it can overcome the shortcomings of the traditional text-based keyword vector space model.

From the perspective of linear algebra, LSA uses the famous matrix theory, with singular value decomposition (SVD) as its key procedure. SVD has important applications in the domain of signal processing and statistical analysis. Given a matrix $\boldsymbol{A}$, SVD can be used to reduce the dimensions of its rows and columns to $K \times K$. We first present the relevant theorems and the definitions.

Definition 1 (singular value) A non-negative real number σ is a singular value for $\boldsymbol{A}$ if and only if there exist unit-length vectors $\boldsymbol{u}$ in k^m and $\boldsymbol{v}$ in k^n such that: $\boldsymbol{vAv} = \sigma\boldsymbol{u}$ and $\boldsymbol{Au} = \sigma\boldsymbol{v}$. The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are called left-singular and right-singular vectors for σ, respectively.

Theorem 1 (SVD) Suppose $\boldsymbol{A}$ is an $m \times n$ matrix, then there exists a factorization of the form $\boldsymbol{A} = \boldsymbol{TSD}^{\mathrm{T}}$, where $\boldsymbol{T}$ is an m-order orthogonal matrix and $\boldsymbol{D}$ is an n-order orthogonal matrix.

Theorem 2 Every matrix has a singular value decomposition.

3 Application of LSA to Query Interfaces

In text retrieval, the three matrices $\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{D}$ that we obtain after SVD have clearly quite physical meanings. Matrix $\boldsymbol{T}$ is a word-to-word relational matrix, denoting the relevance between words. And matrix $\boldsymbol{D}$ is a relational matrix concerning

the documents. The middle matrix $\boldsymbol{S}$ illustrates the relevance between the words and the articles. Applying SVD to the deep web, we construct a mapping between documents and forms. Such a mapping is fairly reasonable. Then, we begin to construct a sparse interface-form matrix.

Suppose that we have collected M forms from M corresponding sites which are represented by a set $F = \{f_1, f_2, \ldots, f_M\}$. Let $\boldsymbol{A} = \{a_1, a_2, \ldots, a_N\}$ be the set of the attributes that we obtain from the query interfaces, and $\boldsymbol{X} = (n(f_i, a_j))_{ij}$ is a matrix with an attribute as a row, a form as a column, where $n(f_i, a_j)$ is the frequency of the attribute j in the form i. For simplicity, if form i contains the attribute, then $n(f_i, a_j)$ is 1; otherwise, it is 0. After the SVD, we can calculate the following similarities:

The similarities between the forms, after the singular value decomposition, are algebraically equivalent to:

$$\hat{\boldsymbol{X}}^{\mathrm{T}}\hat{\boldsymbol{X}} = (\boldsymbol{D}_k\boldsymbol{S}_k)(\boldsymbol{D}_k\boldsymbol{S}_k)^{\mathrm{T}} \tag{1}$$

The similarities between attributes i and j are calculated by the inner-product of the line i and j in matrix $\boldsymbol{TS}$:

$$\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^{\mathrm{T}} = (\boldsymbol{T}_k\boldsymbol{S}_k)(\boldsymbol{T}_k\boldsymbol{S}_k)^{\mathrm{T}} \tag{2}$$

Finally, the similarity between attribute i and form j is measured by the inner-product of line i in $\boldsymbol{TS}^{1/2}$ and line j in $\boldsymbol{DS}^{1/2}$:

$$\hat{\boldsymbol{X}} = (\boldsymbol{TS}^{1/2})(\boldsymbol{DS}^{1/2}) \tag{3}$$

Here, we need to choose a suitable k. It must be large enough, so that it can reveal the potential semantic structure; it must be suitably small enough to avoid bringing too much noise which affects the results.

New deep web sites will be increasingly added to the database. For a new site, after manual or automatic attribute extraction from the form, the new properties and the form have been added to form a matrix. In order to re-calculate the metric matrix, the block matrix is given below:

$$\boldsymbol{X}_{i+a,j+b} = \begin{bmatrix} \boldsymbol{X}_{i,j} & \boldsymbol{A} \\ 0 & \boldsymbol{B} \end{bmatrix}$$

where $\boldsymbol{X}_{i,j}$ is the original matrix; $\boldsymbol{A}$ is an $i \times b$ matrix; $\boldsymbol{B}$ is an $a \times b$ matrix; a is the number of new attributes, and b is the number of the new forms.

From the characteristics of the block matrix, the corresponding matrix results of $\boldsymbol{X}_{i,j}$ are unchanged after decomposition; thus, we can incrementally conduct the singular value decomposition, and obtain the new similarity values.

4 Real Data Experiments

The experimental data we adopt is part of the UIUC web integration repository[12], which is collected manually and stored in the form of xml. First, we extract the attributes from each page of the interfaces with a Java application, and we choose a group of data in the field of "Automobiles", obtaining 210 properties and 97 forms, which forms a 210 × 97 sparse matrix. After processing the attributes form matrix, we obtain the revised relationships between attributes and forms, the similarities between forms and the similarities between the attributes. The first ten sites which are most relevant with site 1(http://www.1stopauto.com) are listed in Tab. 1.

Tab. 1 The ten sites most relevant to site 1

Rank	Index	Source name
1	1	http://www.1stopauto.com/
2	31	http://www.cars.com/
3	92	http://www.stoneage.com/
4	55	http://www.bigjons.com/
5	30	http://www.cars.com/
6	77	http://www.motornet.ie/
7	42	http://www.cybermotors.com/
8	73	http://krieger.smartwebconcepts.com/
9	21	http://www.bigbillybarrett.com/
10	48	http://www.drive.com.au/

5 Conclusion

This paper proposes a method to locate deep web data source based on the theory of LSA to revise the relationships between attributes and forms, the similarities between forms and the similarities between the attributes. It is proved effectively by the experiment on real data. These results can be used to provide users with sites that are the most similar to the site they are visiting, or they can act as a search engine to return appropriate sites according to the attributes the user has specified, just by providing some possible words as attributes. Further work is to make the retrieval more accurate by combining other algorithms to eliminate some ambiguities.

References

[1] Bergman M K. The deep web: surfacing hidden value[EB/OL]. (2001-08)[2008-03-25]. http://www.completeplanet.com/Tutorials/DeepWeb/index.asp.

[2] Zhang Z, He B, Chuan K C. Understanding web query interfaces: best-effort parsing with hidden syntax[C]//*Proceedings of the* 2004 *ACM SIGMOD International Conference on Management of Data*. Paris, France, 2004: 107 − 118.

[3] Zhang Z, He B, Chuan K C. Light-weight domain-based form assistant: querying web databases on the fly[C]//*Proceedings of the* 31*st International Conference on Very Large Data Bases*. Trondheim, Norway: VLDB Endowment, 2005: 97 − 108.

[4] Kabra G, Li C, Chuan K C. Query routing: finding ways in the Maze of the deep web[C]//*Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*. Washington, DC, USA: IEEE Computer Society, 2005: 64 − 73.

[5] Caverlee J, Liu L, Buttler D. Probe, cluster, and discover: focused extraction of QA-Pagelets from the deep web[C]//*Proceedings of the* 20*th International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2004: 103.

[6] Ipeirotis P G, Gravano L, Sahami M. Probe, count, and classify: categorizing hidden web databases[C]//*Proceedings of the* 2001 *ACM SIGMOD International Conference on Management of Data*. Santa Barbara, CA, USA, 2001: 67 − 78.

[7] Raghavan S, Garcia-Molina H. Crawling the hidden web[C]//*Proceedings of the* 27*th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001: 129 − 138.

[8] Ntoulas A, Zerfos P, Cho J. Downloading textual hidden web content through keyword queries[C]//*Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*. Denver, CO, USA, 2005: 100 - 109.
[9] Wang Huili, Liu Wenyu. The latent semantic analysis: rationale and application [J]. *Journal of Huazhong University of Science and Technology*: *Social Science Edition*, 2004, **18** (4): 91 - 94. (in Chinese)
[10] Oates T, Bhat V, Shanbhag V. Using latent semantic analysis to find different names for the same entity in free text[C]// *Proceedings of the 4th International Workshop on Web Information and Data Management*. McLean, Virginia, USA, 2002: 31 - 35.
[11] Li Li, Zhang Taihong, Li Xia. Application of latent semantic analysis to Chinese text classification[J]. *Journal of Xinjiang Agricultural University*, 2006, **29**(2): 99 - 102. (in Chinese)
[12] The UIUC web integration repository [EB/OL]. (2003) [2008-03-15]. http://metaquerier.cs.uiuc.edu/repository.

Deep web 站点查询界面的潜在语义分析

茅琴娇[1] 冯博琴[1] 潘善亮[2]

([1] 西安交通大学计算机科学与技术系,西安 710049)
([2] 宁波大学信息科学与工程学院,宁波 315211)

摘要:为了进一步提高搜索引擎的效率,实现对 deep web 中所蕴含的大量有用信息的检索、索引和定位,引入潜在语义分析理论是一种简单而有效的方法. 通过对作为 deep web 站点入口的查询界面里的表单属性进行潜在语义分析,从表单属性中挖掘出潜在语义结构,并实现一定程度上的降维. 利用这种潜在语义结构,推断对应站点的数据内容并改善不同站点的相似度计算. 实验结果显示,潜在语义分析修正和改善了 deep web 站点的表单属性的语义理解,弥补了单纯的关键字匹配带来的一些不足. 该方法可以被用来实现为某一站点查找网络上相似度高的站点及通过键入表单属性给出拥有相似表单的站点列表.

关键词:deep web;信息检索;潜在语义分析;奇异值分解
中图分类号:TP311