西北大学学报(自然科学版)
2006年 12月,第 36卷第 6期, Dec ,2006, Vol 36, No. 6
Journal of Northwest University (Natural Science Edition)

基于海量数据挖掘的个性化推荐系统

郭 晔[1],王浩鸣[1、2],杨新安[1]

(1.西安财经学院 计算机科学系,陕西 西安 710061; 2.瑞士洛桑联邦理工大学 信息与计算机系,洛桑 1015)

摘要:**目的** 建立海量数据环境中具有个性化的推荐系统。**方法** 在普通文献推荐系统的基础上,增加基于链接页面的 Pagerank计算,从而更精确地表示查询页面相对于特定用户的查询价值。**结果** 结合了基于页面内容的查询方法与基于链接的查询方法的优点。**结论** 具有一定的研究价值,值得在未来的研究工作中加以完善。

关 键 词:文档分类;特征提取;向量空间;邻接矩阵; Pagerank

中图分类号: TP311. 13 **文献标识码**: A **文章编号**: 1000-274 X (2006) 06-0899-04

Web搜索引擎技术中基于 P2P结构的方法近来得到了广泛的研究,其主要是针对基于内容检索的技术。这些技术在如何对基于链接的页面进行计算与排队以便提高检索精度的算法问题上都没有很详细的描述。

本文提出了一种基于智能个性化的信息检索模型,目的是为用户提供特定研究领域的信息检索服务。系统接收用户的请求,根据数据库中用户的特征解释用户的请求,然后向本地数据库与 Internet搜索引擎同时发出经过解释后的请求,对搜索引擎返回的结果计算其重要性,以列表的形式提交给用户,用户感兴趣的内容将被下载到本地,保存在本地数据库中,同时用户的特征数据库将根据用户感兴趣的内容加以维护。

1 基本概念

1.1 TFIDF与文本分类

TFIDF (term frequency/inverse document frequency)常在 VSM (vector space model)中被用来计算文档的权重。对于文本分类领域,权重函数与两个重要的机器学习模型有关: kNN (k-nearest neighbor) 和 SVM (support vector machine)。 TFIDF计算每一个分量(通常是指单词表中出现的词)的权重。

对于特定类型研究领域 c(以下称为"类"),可以通过集合其包含的各个文档的向量 d构成其向量 c。

假设 d表示文档 d的向量, d与各个类向量 c之间的余弦距离,实际表示该文档与特定类的相似程度,其中取值最大的,即为 d所属的类 c。

1.2 Pagerank

对于搜索引擎,在通过查询请求得到相应的检索结果后,需要经过适当计算,从而将查询结果按重要性进行排列。其中一个重要的方法就是利用页面本身所具有的链接, Google是采用这种方式取得成功的例子,它采用了 Brin与 Page[3]发明的链接计算算法。

如果把网页与其链接看作是有向图 G = P(Page, Link),可以用邻接矩阵 P来进行描述, P的元素可以定义为

$$P_{ij} = \begin{cases} 1, \text{Link } i \to j \text{ exists} \\ 0, \text{Otherwise} \end{cases}。$$

其中: $i, j \in (1 \cdots n)$, n表示网页总数。如果从某个网页出发到达其他网页的总概率为 1,则矩阵 P中的元素通过 $P_{ij} = P_{ij}/\deg(i)$ 运算后,可以表示从该页出发到达其他页面的概率。其中, $\deg(i)$ 表示该页面的出度。P称为行随机矩阵 (row-stochastic)。如果考虑网页链接的具体特性,可以认为 P实际对应着马尔可夫链。

收稿日期: 2006-05-28

基金项目:陕西省自然科学基金资助项目 (2005F08)

作者简介:郭晔 (1961—),女,西安财经学院副教授,从事海量数据环境中的信息检索,数据挖掘方面的研究。

修改矩阵 P,对矩阵中全 0的行元素,替换为 $v=\left(\frac{1}{n}\right)e^T$,其中:$n$为页面总数;$e^T$为全 1的行向量。P 被修改为

$$P' = P + d \cdot v^T。$$

其中:$d=\begin{cases}1, & \text{if } \deg(i)=0\\ 0, & \text{otherwise}\end{cases}$

称为 dangling页面指示器。可以验证 P 为行随机矩阵[5]。

P 对应着 G上的随机转移矩阵,因此, Pagerank 可以定义为下列递归过程的极限值

$$x_j^{(k+1)} = \sum_i P'_{ij} x_i^{(k)} = \sum_{i\to j} x_i^{(k)}/\deg(i)。$$

上式可以转化为求 P 的固有特征向量。由于 P′ 中包含有值为 0的元素,这就不能保证 P 具有正的、最大固有特征向量。其原因是 P 可能是可约的,即 P 可能是由几个相互不联通的子图构成。

$Q = cP' + (1-c)ev^T$, $e = (1, 1, 1, \cdots, 1)^T$。

$c \in (0, 1)$,在大多数的文献中,设定为 $c \in [0.85, 1)$, c被称为 dangling系数。

经过处理后的 Q,可以验证 $Q_{ii}^{(k)} > 0$ $(i, k \in (1\cdots n))$,因此 Q是非周期的。可以认为 P 是具有不可约、非周期的正的行随机矩阵。根据 Perron-Frobenius定理,方程式:$x^{(k+1)} = Q^T x^{(k)}$ 一定收敛,并且具有一个最大的、正的固有特征向量。

2 系统结构描述

系统结构如图 1所示。

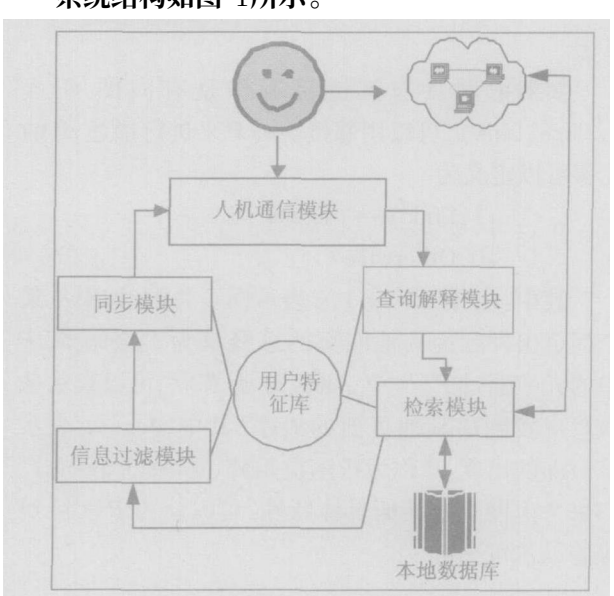

图 1 系统结构图

Fig 1 Architecture of the system

系统由 5个模块、1个用户特征数据库与 1个本地数据库构成,其中:

1) 人机通信 (MM) 模块:用户通过该模块输入查询关键字。最后的查询结果也将通过这个界面显示给用户。

2) 查询解释 (QI) 模块:用作扩展用户的查询要求。对于不同研究领域的用户来说,并不是所有用户都能很准确地表达自己的查询需求。因此,系统将根据用户数据库中用户的特征,对用户输入的关键字进行解释、扩展,以求更准确地表达用户的需求,同时也将查询结果以合适的顺序提交给用户。

3) 查询 (RE) 模块:用作向 Internet以及本地数据库发送查询请求。用户在进行信息检索时有两方面的要求:查准率与查全率。从 Internet上返回的信息可能涵盖更大的检索范围,但从本地数据库中检索出的数据可能对特定用户来说具有更高的查准率。

4) 信息过滤 (FI) 模块:从 Internet与本地数据库中得到的结果是个粗糙的集合,需要根据用户数据库中的特征量与检索结果进行比较,将检索结果重新排序,相关度较高的排在前面。

5) 分析与同步 (AS) 模块:用作下载用户浏览过的网页,根据要求计算各网页的 pagerank值并保存到本地数据库中。同时,根据计算结果维护用户特征数据库。

6) 用户特征数据库:①所有用户的查询历史。记录用户在过去一段时间内的查询活动,其目的是根据这些活动"推测"用户感兴趣的研究领域;②根据用户的查询历史计算得到的用户特征值。通过计算用户访问的页面计算每个页面的特征值,判断这些页面所属的领域,构成用户的兴趣特征。

7) 本地数据库:用于保存从 Internet根据用户浏览历史下载的页面。这些页面与传统的关系数据库中存放的数据有所差别。页面数据大多是非结构或半结构化的,需要经过转化才能保存到关系型数据库中。

3 网页价值重新计算

3.1 关键词提取

在本文涉及的系统中,需要提取特定领域的关键词。

文献[4]介绍了一种新的用于特殊分类的基于统计的词权重算法。这种算法保证了一些无关的词具有较小的权重,从而保证特征词的固有特性在分

类过程中可以得到充分体现。

本文采用称为选取权重 top-n关键词的方法,这些词以下称为类关键词。选取文档源是经过手工分类过的各个领域文档,这些关键词都经过权重计算。假设特定类用向量 $D_i = \{(k_j, w_j), j \in (1, m)\}, i \in (1, n)$ 表示。其中:n表示类的总数,m 表示特定类中类关键词的数目;(k_j, w_j) 表示特定的关键词与其权重。

类似地,用向量 d_i表示文档 d_i: $d_i = \{(k_j, w_j), j \in (1, m)\}, i \in (1, n)$。其中:$n$表示文档的数目,$m$ 是特定文档中关键词的数目;(k_j, w_j) 表示关键词在特定文档中的权重。

3.2　下载链接

根据服务器的登录日志,可以得到用户的访问序列,通过该序列可以从 Internet下载用户访问过的页面。

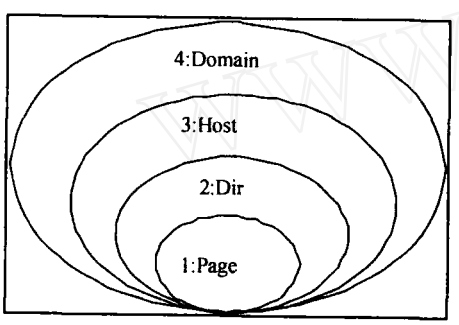

图 2　网站信息组织层次图

Fig 2　Arrangement of information in website

在图 2中,将信息在网站上的组织分为 4个层次:①页面层;②目录层;如: http://liawww.epfl.ch/Research; ③主机层; 如: http://liawww.epfl.ch; ④域层;如 http://www.epfl.ch。

文献[6]中指出:为了得到更好的性能,对于信息结构的分块应该放在第 3层,即主机层。因此,本系统将信息的下载定位于第 4层,即下载第 4层以下的所有页面。

3.3　建立页面链接矩阵

在第 1.2节中,详细介绍了网页 Pagerank算法,那也是 Google赖以成功的基础。

文献[2, 5, 6]提出了一些改进计算效果的有效方法。

在本系统中,页面链接可以分为两种情况:主机内链接(下称内链接)与主机间链接(下称外链接)。这两种链接应该占有不同的份量,设内链接权重为 3/4,主机间链接权重 1/4。则链接矩阵 P为

$$P_{ij} = \begin{cases} \dfrac{3}{4 * \deg(\text{intra})}, & i, j \in H_m \\ \dfrac{1}{4 * \deg(\text{inter})}, & i \in H_m, i \in H_n, m \neq n \\ 0。\text{Otherwise} \end{cases}$$

矩阵 P应该按照 1.2节中所述方法加以处理,从而保证 $P^{(k)}$ 有最大、正特征值。

3.4　Pagerank计算

在本地数据库中,保存有以下数据:①根据登录日志从网站下载的页面;②页面与类的向量表示;③相关页面的链接信息。同时还保存有两个队列:①页面队列,根据页面与类的余弦距离排序;②页面在类中的重要性排序,根据 Pagerank公式计算而得。

对于两个队列,可以设想采取不同的权重,从而判断出特定文档对于特定类的重要程度。

$$\text{score}(d) = \alpha * \text{sim}(d, D) + (1 - \alpha) * PR(d)$$

其中:d是保存本地的页面;$\text{sim}(d, D)$是页面 d 与类 D 的余弦距离;$PR(d)$是页面 d的 Pagerank值;$\alpha \in (0, 1)$称为权重系数。

计算的结果,是新产生一个队列,表示对于特定的类 D重新生成重要性由高到低的页面 d排列。

4　改进内容

在该系统中,使用了传统的算法产生文档与类的特征向量,特征实际表现为出现的"词",实际上对于一些特定的词应该比其他的词具有较高的权重,本系统没有考虑这个情况。

对于 Pagerank的计算,系统模仿了 Google使用的页面价值计算方法,但与它不同的是,Google计算了大范围内的链接情况,而该系统只是集中于同一个"域"层,没有考虑到与其他域的相互影响,严格上说应该是 Google的模拟算法。因此,检索的结果必然与事实之间存在着差距。

对于 Pagerank计算,文献[2]中提出了两个问题:①可能一个页面与特定的类有关联,但没有出现指定的关键词,这就导致该页面没有被检索到;②可能一个网页有很多链接,并且 Pagerank可能也很高(如 Link farm),但与欲查询的内容关联并不是非常紧密。

Google声称已经解决了第 2个问题,但目前很少发现文献介绍如何解决这个问题。

对于 SVM理论,用于处理模式化的数据具有比较高的效率,但 Internet上的数据,特别是对于商业

类的数据,它们的组织与科技类文献迥异,如何有效地处理这些数据,需要新的算法与实验。

5 结 语

本文介绍了传统的搜索引擎不能提供具有个性化的服务原因,试图在推荐技术与 Pagerank计算的基础上,为用户提供具有个性化的服务。

系统首先通过已经分类的文档,选取特定类的特征值,组成特征向量;当用户登录系统,并提交查询关键字后,将计算关键字与各个类的距离,将相关联各类中与查询关键字最相近的页面选择出来;再根据提交给搜索引擎后返回的结果,一起合成输出序列提交给用户。

用户登录后,系统下载用户浏览的页面与链接,利用链接计算出各个页面的 Pagerank值,并将该信息保存到数据库中,同时更新用户特征数据库。

该系统在传统的文献推荐系统的基础上,采用基于链接的页面重要度计算方法,同时保存用户的特征数据,进一步改善了检索精度,适合对特定研究领域有较高需求的用户。

参考文献:

[1] RAFIEI D, MENDELZON A. What is this page known for? computing Web page reputations[C]// HERMAN I The International Journal of Computer and Telecommunications Networking Amsterdam: North-Holland Publishing Co, 2000: 823-835.

[2] FANG Hui, ZHAI Cheng-xiang. Semantics: Semantic term matching in axiomatic approaches to information retrieval[C]// EFTHIMIADIS E N. Proc of 29th ACM SIGIR conference. Karlsruhe: ACM Press, 2006: 115-122.

[3] LEMPEL R, MORAN S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect[C]// HERMAN I Proc of 9th WWW conference. Amsterdam: North-Holland Publishing Co, 2000: 387-401.

[4] XING Wen-pu, GHORBANI A. Weighted pagerank algorithm[C]// KAELING L P. Proc of 2th Annual Conference on Communication Networks and Service Research. New York: IEEE Ycomputer Society, 2004: 305-314.

[5] SOUCY P, MINEAU G W, BEYOND T. Weighting for text categorization in the vector space model[C]. International Joint Conference on Artificial Intelligence. Heidelberg: Springer, 2005: 1 130-1 135.

[6] JIANG Xue-mei, XUE Gui-rong, SONG Wen-guan, et al Exploiting PageRank at Different Block Level[C]// ZHOU Xiao-fang Proc of WISE. Heidelberg: Springer, 2004: 241-252.

[7] 张彤,张[illegible],李军怀. 基于 XML的 Web信息系统中数据访问性能优化方法[J]. 西北大学学报:自然科学版, 2006, 36(3): 398-402.

(编 辑 曹大刚)

Research of personalized recommender system based on data mining on magnanimity data

GUO Ye[1], WANG Hao-ming[1,2], YANG Xin-an[1]

(1. School of Information, Xi'an University of Finance & Economics, Xi'an, 710061, China; 2. School of Information & Computer, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland)

Abstract: Aim To setup a personalized recommender system based on on data mining on magnanimity data. **Methods** Present a personalized recommender system model combining the text categorization with the pagerank. The features of pages were extracted in order to form the feature vector, which will be used in computing the difference between the documents or keywords with the user's interests and the given domain. The links between the pages were divided into two parts, the inter-link and the intra-link, according to the position of pages. All links were of different weight in the link matrix. The final order of the documents was determined by the vector distance and the eigenvector of the link matrix. **Results** It combined the advantages of the method based on content and on links. **Conclusion** It is valuable to be researched in the future.

Key words: text categorization; feature extraction; vector distance; link matrix; pagerank