第 32 卷 第 8 期 计 算 机 工 程 2006 年 4 月
Vol.32 № 8 Computer Engineering April 2006
·博士论文· 文章编号：1000—3428(2006)08—0045—03 文献标识码：A 中图分类号：TP393

基于 P2P 网络的信息过滤与推荐技术研究

李绍滋[1,2]，周昌乐[2]，陈火旺[1]
(1. 国防科学技术大学计算机学院，长沙 410073；2. 厦门大学计算机科学系，厦门 361005)

摘 要：共享信息的集中存储对存放这些信息的服务器提出了较高的要求，同时，服务器将成为整个系统的瓶颈。为此，提出了一种基于 P2P 的信息共享与推荐模型，解决了信息集中存放产生的问题。接着，对该模型中涉及到的基于内容的过滤，提出了一种基于词汇链的方法，较好地解决了纯粹单一关键词无法准确描述文本的问题，并对信息推荐中使用最成功的协同过滤算法进行了描述。给出了文本过滤的实验结果及其分析。

关键词：对等网络；客户机/服务器；词汇链；文本过滤；协同过滤

Research on Information Filtering and Recommendation Based on Peer to Peer Network

LI Shaozi[1,2], ZHOU Changle[2], CHEN Huowang[1]
(1．School of Computer, National University of Defense Technology, Changsha 410073;
2．Department of Computer Science, Xiamen University, Xiamen 361005)

【Abstract】 To solve the bottleneck of the server and the shortage of reliability about centralizing storage in sharing information system, the distributed information sharing model is put forward, which is based on peer to peer networking. Based on it, the basic theory and the algorithm about content-based documents filtering based on lexical chains are given, and then, the collaborative filtering algorithm is discussed. Finally, the validity of content-based documents filtering algorithm is validated through using the medical corpus OHSUMED on TREC-9.

【Key words】 Peer to Peer networking; Client/server; Lexical chain; Document filtering; Collaborative filtering

随着互联网的高速发展和广泛应用，互联网不仅提供了巨大的信息资源，而且提供了一个便利的信息交流与共享平台。但巨大的信息又使我们获取有用信息感到困难，同时，如何充分利用互联网这一便利的信息交流平台合作完成一项科研任务，充分共享合作者所获得的信息，是摆在我们面前的问题。

基于内容的信息检索和信息过滤技术，可以使我们有效获取感兴趣的信息，协同过滤或信息推荐为具有相同或相近兴趣者高效地共享信息提供了可能。

1 基于 P2P 网络的混合过滤模型

根据 P2P 网络特点，提出一个基于 P2P 网络的信息共享和分布存储模型，考虑到纯 P2P 网络的管理较为复杂，这里采用混合 P2P 网络结构。也就是说，真正的文本信息分布存储到每个用户的机器中，系统中另外建立一个服务器，该服务器不存储任何共享信息，仅仅用来存放每个用户共享信息的目录索引(也称为目录服务器)。

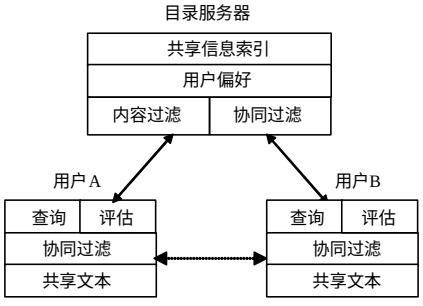

图 1 基于 P2P 的信息共享与推荐模型

当然，每个客户机还可以利用其中的协同过滤系统，向服务器推荐提供自己机器上新获得的信息。整个信息共享与推荐模型如图 1 所示。

2 基于词汇链的文本过滤模型

2.1 构建词汇链的方法

构建词汇链的方法参见文献[1,2]，这里对其作简要概述。在构造词汇链表示文本时，首先要考虑词汇的选择问题，就是哪些词汇适合于作为候选词。在经过文本分析之后，我们删除文本中的所有虚词包括代词、情态动词、介词或带从句的副词、冠词。还有一些出现频率比较高的词，比如：good，do，taking 等，也把它们放入停用词表中。剩下的词汇都把它们列入候选词。

接下来需考虑的是词汇关系问题，在构建时，本文是以词典中的词汇关系来自动完成的。在知识来源方面，我们利用了 WordNet。WordNet 用大家熟悉的拼法来表示词形，用同义词集 Synsets(在一定上下文中可以互换的同义词形的列表)来表示词义。WordNet 目前包括了大约 95 600 个词条，其中包括单纯词 51 500 个，复合词 44 100 个，它们被组织成约

基金项目：国家“863”计划基金资助项目(2001AA114110)；福建省自然科学基金资助项目(A0310009)；福建省科技计划基金资助项目(2001J005)；厦门大学“985”二期信息创新平台项目和厦门大学院士启动基金资助项目

作者简介：李绍滋(1963—)，男，教授、博士生，研究方向：信息检索与过滤，网络多媒体与 CSCW 技术；周昌乐，教授、博导；陈火旺，教授、博导、中国工程院院士

收稿日期：2005-06-01 **E-mail：**szlig@xmu.edu.cn

70 100 个词义或同义词集，描写了上下位、同义、反义、部分-整体等词汇语义关系。

这里主要考虑 3 种词汇关联关系：超强关联、强关联，中等关联。

2.2 文本表示方法的改进

传统的信息过滤技术主要是采用关键词查找和统计技术来过滤相关的文本，只是分别计算各个关键词在文本中出现的频率，而忽略关键词之间的语义关系，忽略主题与关键词之间的关系，对于这种把关键词孤立起来的技术来说具有很大的局限性。这里分析传统关键词技术不足的 3 种主要因素及给出相应的解决方法：

(1)无用词的影响

各个类别中均可以出现的特征，它不代表类别的特点。这些词有些是属于停用词，还有就是那些信息量不大的词，对于这些词将其删除。

(2)词间关系的影响

可分两种情况说明：一种是同义词的影响(比如计算机与电脑)；另一种是具有某种语义关联词的影响，比如：医疗类中，“ 医生 ”、“ 护士 ”、“ 医院 ”、“ 病床 ”、“ 手术室 ”、“ 诊断 ”、“ 药方 ”、“ 感染 ”、“ 病情 ”、“ 抗体 ” 等词是存在某种关联的。其中一个特征的存在某种程度上具有替代其它词的作用，各个特征单独出现的频率可能比较小，而且也许会被一些无关的出现频率大的词所覆盖，利用上述距离公式计算时却没用考虑这种影响，所以同样会导致距离计算不准确。对于这种情况，可以考虑词之间的语义关联，如果这些词共同表达的是一个主题，那么它们就会在词典中的语义距离是比较近的，在文本分析过程中就可以自动地放在一起，而在计算文本相似度的时候就可以把它们综合起来考虑。

比如从文本中抽出这样一些词信息：{{information:3,technique:1,Bayesian-technique:1,datum:2,model:1, area:1}{computer:4}} 其中每个词后面的数字表示在文本中出现的次数。

如果只是分别考虑各个词的词频的话，则 computer 最高，可以知道前面几个词之间有很强的语义关联，它们可以相互补充，从而提高该部分各个词的重要性。

(3)词间地位不平等性的影响

关键词对主题支持作用的大小尽管可用出现次数的加权值大小来体现，但还不够。看文章时并不一定要阅读全文，常常在读了标题或第一段后就可以较准确地确定主题。这说明，存在一些特征词对某一主题具有强的支持作用(决策特征)，它们的存在可以在很大程度上决定文本的主题。而在向量空间模型中，这种决策作用将可能被众多非决策特征的影响所淹没掉。对于这种情况，我们引入特征区域概念。

文本特征区域是能够体现文本主题的区域，包括标题、摘要、关键词、参考文献。但并不是所有的文本主题都有摘要、关键词和参考文献，因此这些结构单元作为可选的单元。国内有人抽样统计，国内中文期刊自然科学论文的标题与文本的基本符合率为 98%，新闻文本的标题与主题的基本符合率为 91%。任何文章几乎都有标题，因此标题是主要的文本特征之一。

线索词是那些总结性或是概括性的标志性词语。比如“ 总之 ”,“ 总而言之 ”,“ 综上所述 ” 等。我们将加强特征区域所包含词的重要性，同样，对于线索词之后的关键词，也将给予增加权重，从而突出该词的重要性。

基于以上的认识，将采用词汇链来表示用户模板和未知文本。首先把用户提交的文本进行分析，然后构造词汇链来表示文本，从而构建出用户模板，来表达用户的需求，这个需求在过滤的过程中采用反馈进行自动学习，使其更精确地满足用户要求。对未知文本，也采用同样方法构建出词汇链来表示文本。

2.3 文本分析

并非文件中所有的词都用于构造文本的词汇链，只有那些最能代表文件所要表达意思的词，也就是关键词汇可被用来构造。文本表示方法可归纳如下：

(1)文本预处理：对词根进行抽取，短语的识别等。

(2)词性标注：对文本中的单词进行词性标注。

(3)关键词抽取：在文本中去除所有属于下列的单词：冠词(如 a, the, an)，介词或连接主句和从句的副词(如 in ,to ,of)，情态动词(如 would,must)和连接词(如 and)等，我们用 W(s,w,c)表示。其中 w 表示这个词，s 表示这个词在此文本中的序号，c 表示该词的词性。比如，(12,think,verb)表示在 W 中第 12 个词是 think，这是一个动词。也可以给各种词性的词赋予不同的权值来表示它们不同的重要性，一般而言，名词要赋以最大的权值。对于那些在标题、首段、末段、段首、段尾出现的词语也可以增加其权重。也可以设一个阈值，把那些出现频率低于该频率的词去除。

(4)词汇链表示文本：在经过了关键词提取之后，得到文本的词的系列，再经过词汇链的自动构建方法，得到文本的词汇链表示。词汇链中的各个关键词的初始权重为该词出现在文本中的频率、词性加权系数、关键词之间的关联权重及关键词出现在特征区域的权值共同组成。

2.4 文本过滤

到此为止，文本和用户的需求已表示成词汇链的形式，文本与用户需求的相关度可以通过以下的形式计算的余弦值来衡量：

$$\cos(a) = \frac{\sum_{ij} V_i * T_j}{\sqrt{\sum_i V_i^2 + \sum_i T_i^2}} \qquad (1)$$

在所需过滤的所有文本当中，可以根据这个值来进行相关度排序反馈给用户，也可以设定一个阈值 k，当某文本与用户需求的相关度大于 k 时则认为该文本符合用户需求，把文本按相关度大小的顺序返回给用户，把低于该值的所有文本去除或存在某处以备用户在有空时处理。可以把用户的反馈考虑进去，若用户认为几乎所有所过滤出的文件都是他所感兴趣的，则我们可调低 k 值，相反，若有很多文本不符合用户的兴趣，则调高 k 值。

3 协同过滤算法

实现信息推荐有多种方法[3]，协同过滤技术是实现信息推荐技术的最成功方法。协同过滤的主要思想是根据某个用户与其他用户兴趣的相似程度，决定该用户是否将自己的文本推荐给其他用户。如果两用户的兴趣爱好相同，系统将给他们推荐相同的文本信息。因此，协同过滤使用户花较小的代价就能获得有用的信息。

一般来说，协同过滤包括 3 个步骤：(1)寻找相关用户(即近邻)；(2)计算预测的等级；(3)利用预测值进行推荐[4]。

为了产生用户的预测，首先应该标识用户的近邻，近邻用户就是与本用户的兴趣度相关的用户。衡量用户的相关程度通常采用计算二者之间的相关系数的办法。有多种计算相关系数的方法，最常用的方法就是限定皮尔森(Pearson)相关

系数法，为简化，我们采用了如式(2)所示的简化形式：

$$\text{correl}(u_i, u_j) = \frac{u_i^T u_j}{\| u_i \| \| u_j \|} \quad (2)$$

在信息检索中，式(3)也被称为余弦相似度测量。

步骤（2）时，系统将为用户生成预测值。通过每个用户的平均评价等级以及相关联值来计算用户的预测值，即式(3)所示。

$$r_{i,j} = \frac{1}{|N(u_i)|} \sum_{u_k \in N(u_i)} correl(u_i, u_k) * r_{k,j} \quad (3)$$

利用式(3)，可以预测用户的评价等级。有较多的评价等级信息，就可以在下一步为用户推荐较合适的信息。

第(3)步就是预测的应用，包括将上述的 Z 值加回预测值，如果结果较大，将此信息作为推荐信息展示给用户；同时，用户可以对推荐的信息提供评价，并将评价反馈给系统，这样，系统将会从用户的反馈信息中学习到该用户的准确兴趣，以便修正系统的推荐行为。

4 过滤模型的实验结果及实验分析

评价文本过滤系统的两个主要指标是查准率和查全率，本文采用 TREC-9 上的医学语料库 OHSUMED，这是著名的国家医学图书馆的 MEDLINE 医学文献库的一个子集，由 1998~1991 年的医学文献组成，共含有文本 348 566 篇，来自 270 种医学期刊，总容量为 400MB。其中 1987 年的文摘将作为训练语料，而 1988~1991 年的文摘将作为测试语料[5]，测试结果如表 1 所示。

表 1 测试结果

	传统基于关键字的过滤系统	基于词汇链的过滤系统	差距
平均精度	38.53%	47.46%	8.93%

分析上述结果表明：在传统的向量空间模型中，只是单独考虑各个关键词，利用该词出现在文本的次数，而没有充分地利用文本中各词之间的关联信息，而我们采用词汇链的形式，则更加准确地表达了文本主题内容，从而提高过滤的精确度。

5 结论

本文在分析了 P2P 网络优势后，提出了基于 P2P 网络的信息共享模型，针对模型中涉及的基于内容的信息过滤和协同过滤方法进行了较为详细的分析，最后，给出了基于内容的信息过滤算法的实验结果。结果表明：基于词汇链的方法的确优于基于关键词的方法。

参考文献

1 Li Shaozi, You Wenjian. Lexical-chain and Its Application in Text Filtering[C]. Proceedings of the IEEE International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, 2004-04.

2 Silber G, McCoy K. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization[J]. Computational Linguistics, 2003: 29 (1).

3 Terveen L, Hill W. Human Computer Collaboration in Recommender Systems[M]. New York: Addison-Wesley, 2001.

4 Soboroff I, Nicholas C. Collaborative Filtering and the Generalized Vector Space Model[C]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 2000-06: 351-353.

5 Roberton S, Hull D. The TREC-9 Filtering Track Final Report[C]. Proceedings of the Ninth Text Retrieval Conference, 2001-02.

(上接第 13 页)

4.1 配准模版的选取

SPM 方法中，配准模版选用的是第 1 组图像，也就是 04 号图像。这里采用平均图像作为模版。

4.2 图像的配准

Legendre 矩是基于二值图像的配准。对于脑功能的灰度图像，首先提取图像的边界；再对得到的图像边界进行拟合，从而在其上进行基于 Legendre 矩的图像配准。图像的几何变换采用了 3 参数的刚体变换。下面给出了图像的模板(图 3)，待配准的图像(图 4)和最终的配准结果(图 5、图 6)。

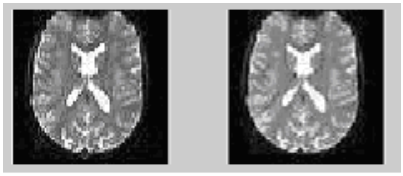

图 3 模板图像 0438　　图 4 待配准图像 2fm2438

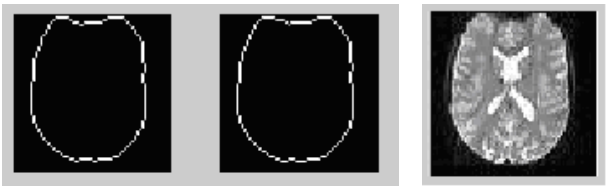

图 5 对图 3、图 4 的二值化处理和曲线拟合　图 6 已经配准的图像

5 结论

基于矩的配准是一种快速图像粗略配准算法，可以在精确图像配准之前进行一个粗配准，从而大大提高精确配准的速度。Legendre 矩是正交矩，本文对其进行简化，将链码和改进的矢量斜率法应用到 Legendre 矩的计算中，在精度相当的情况下，大大简化了目标函数(式(8))的计算复杂度，从而提高了计算的速度。

参考文献

1 Sansone G. Orthogonal Functions[M]. New York: Dove Publications, 1991.

2 葛　云, 舒华忠, 罗立民. 基于 Legendre 正交矩的配准方法及其在二值图像配准中的应用[J]. 电子学报, 2001, 29 (1): 54-56.

3 汪家旺, 舒华忠, 罗立民等. 基于 Legendre 矩的 CT 及 MR 医学图像融合方法[J]. 中国图像图形学报, 2001, 6 (4): 369-373.

4 Jiang X Y, Bunke H. Simple and Fast Computation of Moments[J]. Pattern Recognition, 1991, 24 (8): 801-806.

5 Shu H Z, Luo L M, Yu W X, et al. A New Fast Method for Computing Legendre moments[J]. Pattern Recognition, 2000, 33 (10): 341-348.

6 Arelli C, Massarotti A. On the Parallel Generation of Straight Digital Lines[J]. CGIP, 1978, 7 (1): 67-83.

7 吴立德. 计算机视觉[M]. 上海: 复旦大学出版社, 1993.

8 王明江, 唐璞山. 基于矢量斜率的分段线性拟合[J]. 软件学报, 1999, 10 (2): 165-169.