●刘伟成[1]，孙吉红[2]
（1. 武汉科技大学 管理学院，武汉 430081；2. 武汉大学 信息管理学院，武汉 430072）

基于专题文献的信息内容过滤系统研制与实现

[关键词] 信息过滤系统；专题文献；特征提取；向量空间模型

[摘　要] 针对当前综合性信息过滤系统不能满足不同知识结构、不同查询兴趣的专业人员对特定领域的信息获取的需求，在对经典信息过滤算法进行研究分析的基础上，提出了基于专题文献的信息内容过滤系统的设计，并加以实现。本系统在专业词汇、特征提取、特征加权等方面进行了改进，最后在小范围内进行了测试。

[中图分类号] TP18；TP391.3
[文献标志码] A
[文章编号] 1005－8214(2009)07－0065－04

1 引言

随着网络的发展，人们对信息质量的要求不断提高，面对网上大量的信息，已往的搜索和过滤系统无论是在效率上还是返回结果效果上都面临着前所未有的难题。国内外已有的网络信息过滤系统存在严重的失真现象。目前，国内还没有原创性的完善的专题性搜索引擎。本项目旨在根据主题性搜索引擎的核心模块之——专题性信息过滤模块，来探讨解决这一问题的办法。

一般专业人员使用的测试系统，其目标是有效地解决信息过载问题。由于专业搜索引擎面向某一个领域，能把资源集中并且有效地利用，以最大限度地满足本专业的信息需求，从而能够针对用户的需求和反馈向用户推出个性化的文档。该项目是基于内容的，通过比较信息资源和用户的兴趣模型之间的相似度来向用户推出过滤结果。在用户兴趣采集中，系统可实现用户采用主题词和文档示例的方式；在过滤结果的给出过程中，系统可实现结果的排序，为用户提供更准确的过滤结果。

专题式信息过滤技术不求包罗各个学科、各种类型的信息，但求本专业、本学科的信息最全，采用的技术更具有针对性。因此，其过滤结果更精确，相关性更高，特别适合于不同知识结构、不同查询兴趣的专业用户群体。

2 专题文献信息过滤系统的设计

2.1 设计思想

传统的基于内容的信息过滤系统针对的是普通用户，所过滤的内容一般是综合性的。许多研究机构研制的一些基于用户个性化的信息过滤系统[1]中使用的分词字典大都是采用通用字典，检索的准确度和效率均不够高。本项目试图为特定领域专业人员设计一个专题文献过滤系统：选择某领域中的专业词汇作为分词词典（去掉不常用词汇），提高分词的准确性；降低用户模板和文档模板的维度，提高系统的检索效率。

所选择的专题文献的一般格式包含题名、关键词、摘要等。它们基本上概括了原文的主要内容。针对这种情况，本文从构建高效的特征词分词词典、适当降低文本向量表示的维数、减少文本匹配规模等方面对系统进行了改进。

2.2 系统逻辑结构图

本系统由测试模板模块（包括分词和特征项提取、特征项权重的计算）、用户模板模块（用户模板初始化和用户模板重构）、用户反馈模块、人机界面模块（用户信息管理、用户兴趣输入、过滤结果推出等）几个模块组成。其中测试模板模块是本系统的核心部分（如图1）。

2.3 系统开发环境

（1）操作系统平台：系统开发环境 Window2000 Server。

（2）开发环境平台：采用 Java 语言进行开发。选用 Java 语言作为开发工具的原因是：作为一种程序设计语言，它简单、面向对象、不依赖于机器的结构，可降低编程难度和工作量，缩短开发周期；可提供并

[基金项目] 国家自然科学基金资助项目（70473067）；湖北省教育厅科学研究计划优秀中青年人才项目（Q20081108）

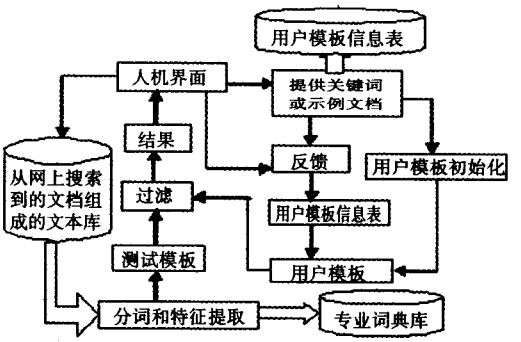

图 1　基于专题的信息过滤系统逻辑模块

发的程序执行机制，具有很高的运行效率；可提高软件的可靠性、移植性、安全性；可最大限度地利用网络资源；提供了丰富的类库，使程序设计者可以很方便地建立自己的系统等。

（3）数据库平台：数据库采用的是 SQL Server2000 数据库。选用 SQL Server2000 数据库作为数据库平台的原因是：SQL Servers 是大型的关系型数据库，方便对数据库联网调用，并支持多机多人协同工作，对于团体用户，可以促成一个共同的团体目标或群体意识的拟合。

3　系统实现的关键技术

3.1　文档和用户模板的表示

常用的基于统计的信息表示模型包括布尔模型、向量空间模型、[2] 概率模型。[3] 其中，由于向量空间模型能方便地将文本内容表示成计算机可识别的形式，从而将文档和用户需求转换成计算机表示的信息。因此，本系统采用向量空间模型来表示文档和用户模板。[4,5]

向量空间表示法把文挡描述成高维空间中的一个向量，向量的分量为文挡中出现的特征词的权值，而文档集中的特征词按一定的顺序排列便构成了文档空间中的维。文档向量的某一维的权值的形成通常是根据文档中该特征词的出现频率经过归一化处理后得到的。更精确的做法是再将其乘以该特征词的反比文档频数。这样，一篇文档就表示成了欧氏空间中的一个向量。同时，把用户输入的查询也经过同样的处理形成向量，就可以用向量夹角余弦或其他公式来表示文档和用户查询之间的相似度。

即：$D\ (<t_1,\ w_1>,<t_2,\ w_2>,\cdots<t_n,\ w_n>)$　(1)

其中，D 为文档，t_i 为表示文档的向量空间中的一个项（词），w_1 为项 t_i 的权重，$1\leqslant i\leqslant n$。

我们从网上选取真实文本作为示例文本，从示例文本中提取过滤模板。

3.2　分词和特征项的提取

要将文本内容和用户模板表示成计算机能处理的数据形式，首先要对其进行分词以得到相应的词集。虽然说，用文本中出现的所有的词集组成文本的向量最接近文本内容，但如此一来会导致向量维数太大，这必然对时间和计算效率产生非常不利的影响。由于本系统是基于专题的信息过滤系统，其过滤对象是专题文献，所选文献格式包含有题名、关键词、摘要等文献的基本信息。

本系统根据特定领域的专业词汇词典对文档进行分词，并统计分词得到的专业词汇的出现频率。由于专题文献的题名、关键词、摘要等能有效地体现全文核心内容，因此，分词和特征项就在其中提取。这样，大大降低了文本向量表示的维数，并且没有削弱文本内容表征的准确性。对于一篇专题文献而言，文档中的特定领域的专业词汇的数量不是很大，因此从文档中分词得到的专业词汇可以作为文档的特征项。

在传统的增字或减字分词算法 [6] 中，每增加或减少一个字就需要重新匹配字符串，效率较低。因此，本系统采用一种特殊的分词算法——直接匹配法，[7] 省略了传统分词算法中字符串的重复匹配工作，提高了效率。

3.3　特征项加权

在用向量空间模型来表示文档的时候，每一特征都用其对文档的贡献来表示，即该特征在文档中的权重。准确地计算出各个特征的权重是很重要的。比较通用的权重计算方法有布尔权重、特征频度、TF-IDF、墒等。[8] 由于本系统是基于专题的信息过滤系统，为了区别题名、关键词、摘要的特征信息在文献中的不同贡献，本文采用不同的权值系数对它们分别进行加权。根据经验值对文献中标引特征词的位置权值 P_j（j=1,2,3,）做如下设定，标题、摘要和关键词分别为 0.7、0.3、0.5。本文提取的特征项就是一些专业词汇，其在文档中出现的频数越大，表示其相关性越好，但不符合著名的TF-IDF 权重的计算规则，即特征在文档中的权重正比于特征在文档中出现的次数而反比于语料中包含该特征的文档的数目，因此，我们采用了特征频度（即特征项在文档中不同位置出现的频数 f_i）来进行特征项的权重计算。这样特征项的真实权重 W_i 就由两部分位置权 P_j（j=1,2,3,）和特征频度 f_i 组成，表示为

$$w_i=\sum_{j=1}^{3} p_j f_{ij} \qquad (2)$$

位置权值如表1所示。

表1　位置权值

位置	标题(P_1)	关键词(P_2)	摘要(P_3)
位置权	0.7	0.3	0.5

通过位置权和特征频度相综合的方法计算得到的特征项的权重是一个常数，这样一旦文档提供后，其向量空间模板将不再改变，这样就大大降低了计算复杂度，提高了运行效率。

3.4　用户模板重构

由于用户表达信息的不准确性和不完全性，系统一次过滤的结果往往达不到用户的要求，因此系统需要根据用户浏览过滤结果的行为和用户提供的反馈信息，及时地更新用户的兴趣。对向量空间模型来说，用户兴趣自适应学习更新的经典方法是Rocchio算法，并且根据不同的需求，产生了很多变种。本文采用文献［9］提供的相关反馈公式

$$q_1=\alpha q_0+\beta\sum_{k=1}^{n_1}\frac{R_k}{n_1}-\gamma\sum_{k=1}^{n_2}\frac{T_k}{n_2} \qquad (3)$$

其中，q_1是新的用户模板，q_0是旧的用户模板，R_k是第k篇相关文档的向量，n_1为相关文档数，T_k是第k篇不相关文档的向量，n_2为不相关文档数。α，β，γ为调节系数，根据经验定义为：$\alpha=0.5$，$\beta=1.0$，$\gamma=0.25$。

3.5　过滤算法

当用户模板构造和文本特征选择完成后，过滤计算就比较简单，主要方法有相似函数和距离函数两类。最常用的就是计算向量间的余弦相似度，用户模板D_1和文档向量D_2之间的相似度可采用向量夹角余弦［10］如下计算：

$$\mathrm{sim}(D_1, D_2)=\cos\theta=\frac{\sum_{k=1}^{N} w_{1k}\cdot w_{2k}}{\sqrt{\sum_{k=1}^{N} w_{1k}^2\sum_{k=1}^{n} w_{2k}^2}} \qquad (4)$$

两向量之间夹角越小，其余弦值越大，说明相似程度越大，文档符合过滤需求的可能性增加。设定一个过滤阀值ψ，当$\mathrm{sim}(D_1, D_2)\geq\psi$，说明特征向量$D_2$所对应的内容符合过滤需求，应该向用户推送。

4　实验结果

我们以长江发电站有关专题文献为过滤主题，从武汉大学图书馆下载了水利水电领域的53篇文档和20篇不相关的文档组成一个测试文档集。同时，从武汉大学图书馆下载了水利水电的专业词汇表作为本系统的分词词典和特征词典，这样系统中就省略了通过训练模板来构建特征词典的模块。但本系统的特征词典具有扩充功能，可根据用户给出的主题词和示例文档中的关键词对特征词典进行扩充。

实验结果表明，在保证实时性的前提下，初次过滤精度可达到71.7%。对实验结果进行多次反馈，过滤精度不断提高，如表2所示。

表2　实验结果

反馈次数	0	1	2	3	4	5
过滤精度	71.7%	75.5%	81.1%	84.9%	86.8%	88.7%

5　系统存在的问题与不足

本文所提出的实验系统已基本实现，但还存在着一些问题需要解决，主要表现在以下几个方面：

（1）本系统只能处理word格式的信息，而网上的信息有HTML格式、PS格式、PDF格式等。因此，下一步工作是使系统能够处理PS格式、PDF、HTML等格式的信息。

（2）数据库采用的是SQL Server2000数据库，在这种数据库中，建立训练文档集导入数据时需要人工将一篇篇文档一一导入，加大了系统管理员的工作量。

（3）在系统运行过程中，当用户给出的主题词或示例文档中的关键词不在专业词典中时，系统首先要将这个主题词添加入专业词典中，然后再利用扩充过的词典进行后续的任务。实验中，此时系统运行速度大为降低。

（4）在过滤计算中，我们仅就小样本训练集进行反复实验测试得到过滤阈值，对于大样本训练集，阈值的合理性需要进一步的测试。

［参考文献］

［1］徐小琳，等．信息过滤技术和个性化信息服务［J］．计算机工程与应用，2003，(9)：182－184.

［2］SALTON. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer［M］. New York: Addison-Westey, Reading, Mass, 1989.

［3］PAZZANI M, BILLSUS D. Learning and revising user profiles: the identification of interesting web sites［J］. Machine Learning, 1997, 27 (3): 313－331.

［4］吴立德．大规模文本处理［M］．上海：复旦大

学出版社，1997.
[5] 杨建林．信息检索模型与逻辑理论［J］．情报学报，2000，19（5）：51—519.
[6] 揭春雨，刘源，梁南元．论汉语自动分词方法［J］．中文信息学报，1989，3（1）：1—8.
[7] 张国煊，等．快速书面汉语自动分词系统及其算法设计［J］．计算机研究与发展，1993（1）：61—65.
[8] K Aas，L Eikvil. Text categorisation: A survey［R］. Norsk Regnesentral: Norwegian Computing Center，1999.
[9] Joon Ho Lee. Combining the evidence of different relevance feedback methods for information retrieval［J］. Information Processing and Management: an International Journal，1998，34（6）：681—691.
[10] 杨小平，丁浩，黄都培．基于向量空间模型的中文信息检索技术研究［J］．计算机工程与应用，2003（15）：109—111.

[作者简介] 刘伟成（1971—），男，山东省莱州市人，博士，副教授，图书馆副馆长，主要研究领域为信息检索与信息服务系统；孙吉红（1971—），女，副教授，博士研究生，主要研究领域为信息检索、信息挖掘。

[收稿日期] 2008—12—24　[责任编辑] 王 岗

（上接第64页） 西华，临川人，乾隆十年（1745年）进士，累官内阁学士兼礼部、工部侍郎，事迹具《清史列传》。史贻直即锡侯座师贻谟兄，字儆弦，号铁崖，康熙三十九年（1700年）进士，仕至文渊阁大学士，《清史列传》《清史稿》皆有传。

[注　释]

① 宜丰县置县于孙吴黄武中。后几经废置，至宋太平兴国中，析高安之盐步镇、天德乡、太平乡及上高之太和乡、宣风乡置新昌县，治盐步镇。故新昌别称盐邑。民国二年，复名宜丰，以别于浙江之新昌。

② 见存王锡侯诸书，均自署“豫章（或作瑞州新昌）王锡侯韩伯氏”；《字贯》自序后又钤有“王锡侯印”、“滨洲”印记二方。

③《清人室名别称字号索引》（增补本）（杨廷福、杨同甫编，上海古籍出版社2001年12月第1版）下册第82页，佚其字号，但记其室名“三树堂”。《国朝诗观》牌：“三树堂藏版”。舒宝璋《呕心〈字贯〉，家破身亡——语文学家王锡侯事略》：“王家门　第寒微，故锡侯以‘三树’名室。唐杜牧《题村舍》诗云：‘三树稚桑春未到，扶床乳女午啼饥。潜销暗铄归何处？万指侯家自不如！’”（载《江西社会科学》1995年第2期）。又《唐诗试帖（课蒙）详解》卷首王锡侯自序，署“乾隆戊寅岁夏五廿日新昌王锡侯书于滋畹堂”。“滋畹堂”殆亦其斋名。

④《宜丰县志》卷四十二《人物》，又卷四十三《附录·王锡侯文字狱纪略》。

⑤ 苏轼《三槐堂铭并叙》，见《全宋文》卷一九八五。

⑥ 乾隆四十二年十月癸丑，谕军机大臣曰：“书中所有参阅姓氏，自系出赀帮助镌之人，概可免其深究。”见《清实录》，中华书局1986年11月第1版。

[参考文献]

[1]（清）王锡侯．《字贯》［M］．清乾隆四十年（1775年）刻，国际文化出版公司出版《字典汇编》本.
[2] 孟森．字贯案［C］// 清代文字狱史料汇编第14册．北京：北京图书馆出版社，2007.
[3] 胡思敬．《国朝瑞新》两学题名录［M］．民国十年（1921年）尊孔会排印本.
[4] 朱汝珍辑．词林辑略［J］// 周骏富辑．清代传记丛刊第16册．台北：明文书局印本.
[5] 冯煦，等．金坛县志［M］．民国十年（1921）刊本.
[6]（清）杨文峰，万廷兰．新昌县志［M］．清乾隆五十八年（1740年）刻增修本.
[7]（清）李景峰，史炳．溧阳县志［M］．清嘉庆十八年（1813年）修光绪二十二年（1896年）重刻本.
[8]（清）法式善．清秘述闻［M］．北京：中华书局，1982.
[9]（清）钱陈群．香树斋文集［M］// 四库未收辑刊玖辑第19册．北京：北京出版社，2000.
[10]（清）王锡侯．唐诗试帖（课蒙）详解［M］．清乾隆二十三年（1758年）滋畹堂刻本.
[11]（清）陈洪书，等．望都县志［M］．清乾隆三十六年（1771年）刻本.
[12]《清实录》［M］．北京：中华书局，1986.

[作者简介] 何巧云(1959-)，女，本科毕业，现为江西省图书馆副研究馆员。

[收稿日期] 2009—02—26　[责任编辑] 张京生