第34卷 第24期 计 算 机 工 程 2008年12月
Vol.34 No.24 Computer Engineering December 2008
·软件技术与数据库· 文章编号：1000—3428(2008)24—0064—03 文献标识码：A 中图分类号：TP311

一种基于内容的混合模式过滤模型

张玲达，金 林，程秀霞，江 飞
(安徽理工大学电气与信息工程学院，淮南 232001)

摘 要：目前的文本内容过滤系统大多是基于关键词的，在对准确性过滤要求不高的情况下可以完成过滤任务。为进一步提高过滤效率，该文提出一种基于内容的混合模式过滤模型，引入语义分析技术，在关键词匹配技术的基础上进行语义框架的匹配，从而保证信息过滤的速度，改善信息过滤的准确度。通过实例对其有效性进行了验证。

关键词：内容过滤；混合模式；语义框架

Composite Mode Filter Model Based on Content

ZHANG Ling-da, JIN Lin, CHENG Xiu-xia, JIANG Fei
(School of Electric and Information Engineering, Anhui Uinversity of Science and Technology, Huainan 232001)

【Abstract】Most text filtering systems are based on keywords, which can accomplish filtering task without high precision requirement. To improve the efficiency of filtering, this paper presents a composite mode filtering model based on content. The matching of semantic framework based on the matching technology of keywords ensures the speed of information filtering, and improves the precision of information filtering. Its validity is proved by instances.

【Key words】content filter; composite mode; semantic framework

1 概述

从1992年起，NIST与DARPA联合赞助了每年一次的文本检索会议(Text Retrieval Conference, TREC)，为文本过滤技术的发展提供了强有力的支持[1]。过去的研究工作大多专注于文本内容的过滤技术，尤其是关键词匹配技术。在提高信息过滤的速度上基于关键词的匹配技术有着自身的优势，实现了网络信息过滤实时性的要求。随着网络用户对安全性要求的不断提高，过滤的准确性显得越来越重要，但关键词匹配技术很难满足这一要求，基于语义分析的过滤技术被提了出来，并在短时间内得到了飞速发展，实现了网络信息过滤准确性的要求，但由于其复杂度过高，因此影响了过滤的速度[2]。

本文设计了一种基于内容的混合模式过滤模型，构造了基于内容的混合模式过滤算法，即在关键词匹配的基础上进行语义框架的匹配，算法的设置不仅满足了网络传输实时性的要求，同时也满足了过滤准确性的要求。

2 基于内容的混合模式过滤模型

文本内容过滤模型的工作流程主要包括以下4个部分：(1)建立用户需求模板；(2)提取待过滤文本的特征向量；(3)用户需求模板与待过滤文本的匹配；(4)文本过滤信息的反馈，进一步改善用户的需求模板。

本文提出的基于内容的混合模式过滤模型如图1所示。其中，概念扩充与关键词匹配2个功能模块组成的第1个虚线框表示该过滤模型所进行的第1步预处理操作——粗选过滤；而填充语义框架与语义框架2个功能模块组成的第2个虚线框表示了过滤模型中所进行的第2步处理操作——细选过滤。该模型中所构造的基于内容的混合模式过滤算法是以基于关键词的匹配作为整个过滤算法的预处理操作，在此基础上实施基于语义框架的过滤匹配，这样的算法设置使得本文提出的混合模式过滤算法既可以保持一定的过滤速度，同时也提高了过滤的准确性，是一个比较适合网络信息安全性过滤的算法[3]。

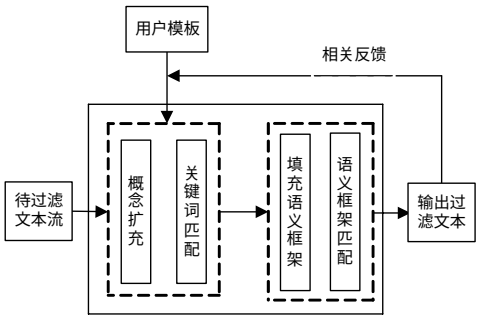

图1 混合模式文本过滤模型

2.1 用户模板的建立

用户需求模板用于揭示一个或一组用户长期的信息需求，由用户建模组件来实现。用户需求模板是个性化信息服务的基础，模板的准确性和实效性直接决定服务质量的优劣，即过滤系统的性能。对于用户的需求模板，既可以从正面揭示，也可以从反面解释，也就是说，既可以描述用户需要的感兴趣的信息，也可以描述用户不需要的不感兴趣的信息。常用的方法是通过抽取与兴趣相关的特征词列表来表达用户需求模板，并借助学习算法，动态调整特征项权值以适应用户兴趣的变化。

基金项目：国家部委“十五”预研基金资助项目
作者简介：张玲达(1982－)，女，硕士研究生，主研方向：智能仪表，信息处理；金 林，高级工程师；程秀霞，硕士研究生；江 飞，硕士
收稿日期：2008-06-22 **E-mail：**zhld1220@163.com

(1)初始用户模板的建立

用户信息需求模型的建立是文本信息过滤的一个基本任务。用户信息需求模型的逻辑表示为用户信息需求模型的外部表示形式，称为用户模板，简称为模板P，最常用的为基于特征项(关键词)的表示方法，记为$U=(t_1, t_2,\cdots, t_n)$，其中，t_i 为特征项，$1\leq i\leq n$，n为用户模板的维数。基于特征项的用户需求模型是用户按照规定的格式将自己感兴趣的关键词以及处理要求发送到匹配规则库，根据用户新的需求过滤待匹配的文本，并按照用户设置的要求提供给用户。

(2)概念扩充后的用户模板

由于自然语言中同义、近义现象普遍存在，因此会出现用户所用的特征词与文本关键词不一致，但实际上两者是匹配的，也就是说，表达相同或相近主题内容的文本，虽然所用的词汇不完全相同，但在语义概念的层次上拥有很多共同项。此外，在同一篇文本中，提取特征时也应该考虑同义词、近义词合并等处理，因此，有必要将词汇映射到概念一级，以便更加全面地反映文本之间的相似度。

本文采用的中文 WordNet 概念词典信息十分丰富，包括词条信息、词法分析和生成的信息、句法分析和生成的信息、语义分析的信息和概念处理信息。在词典中，语义码相同的词为同义词，一组同义词为一个词群。

假设用户提出的关键词集合$U=(t_1, t_2,\cdots, t_n)$作为用户模板P，并假定关键词间的关系为“ AND ”，下面将其映射到概念空间的概念特征集V，定义概念映射如下：

$$U=(t_1, t_2,\cdots,t_n)\xrightarrow{\Phi}(\Omega(t_1),\Omega(t_2),\cdots,\Omega(t_n))=V$$

其中，$\Omega(t_i)=(<c_{i1},w_{i1}>,<c_{i2},w_{i2}>,\cdots,<c_{is_i},w_{is_i}>),\{c_{i1},c_{i2},\cdots,c_{is_i}\}$ 是概念词典中与 t_i 具有相同语义码的概念集合，w_{ij} 是 c_{ij} 的权重。当 $w_{ij}=1$ 时，若 $t_i=c_{ij}$，称 w_{ij} 为 t_i 的原始权重；当 $w_{ij}=\alpha\ (0<\alpha<1)$ 时，若 $t_i\neq c_{ij}$，称 w_{ij} 为 t_i 的扩充权重。如果在概念词典中不存在 t_i，则 $\Omega(t_i)=(<t_i,1>)$。

概念扩充后的用户模板能更全面地反映用户的信息需求，通过加入扩充项，增加了同义匹配的机会，有利于获得包含同义不同形的相关文本，提高了查全率和匹配的效率。

2.2 基于关键词的匹配

基于关键词匹配的预处理操作是基于内容的混合模式过滤模型中的第 1 步匹配过滤，本文的关键词匹配算法应用了特征向量计算法，把文本过滤问题转化成向量余弦的计算问题[4]。利用用户模板U与待过滤文本V的夹角余弦来衡量待过滤文本与用户模板的相似度，通过比较相似度值与初始定义的过滤阈值提取相关的文本，完成第 1 步粗选过滤的任务。

经过整理，设扩充后的概念特征集为$\{x_1,x_2,\cdots, x_m\}$，则基于关键词的用户模板为$U=(u_1, u_2,\cdots, u_m)$，u_i为项在概念扩充中定义的权重。设文本T的特征向量为$\boldsymbol{V}=(v_1,v_2,\cdots, v_m)$，$v_i$为项$x_i$的权重，计算公式为

$$v_i=f(x_i)=\frac{f_t(x_i)\,\mathrm{lb}(1+f_v(x_i))^l}{C\sqrt{\sum_{j=1}^{m}(f_t(x_j)\,\mathrm{lb}(1+f_v(x_j))^l)^2}}(1+\alpha)$$

其中，$f(x_i)$ 表示x_i的权重函数；$f_t(t_i)$ 表示x_i在文本内的频数；$f_v(t_i)$ 表示x_i的段落频率；l 表示主题词t_i的词长；C是比例因子。当主题词位于段首、段尾和结论性句子时，a =0.5，否则a=0。

权重公式设计的基本思想是：关键词的段落频率越高，表明反映文本主题的能力越强，应赋予较大的权重。另外，短词具有较高的频率、更多的含义，是面向功能的；而长词的频率较低，是面向内容的。加大长词的权重、增强词汇的区分度，也可以减轻因分词造成的单个汉字或词的不稳定性。文章的标题与主题紧密相关，每个段落的开头和结尾含有的关键词应给予较高的权重。

将用户模板与待过滤的文本进行相似度运算，即向量内积运算，计算公式为

$$sim(U,V)=\frac{\sum_{i=1}^{m}u_iv_i}{\sqrt{\sum_{i=1}^{m}u_i^{\,2}\sum_{i=1}^{m}v_i^{\,2}}}$$

选定相似度阈值 θ，若 $sim(U,V)>\theta$，则认为文本 V 与模板 U 相匹配。但如此过滤有时显得粗糙一些，可能会产生匹配特征虽然分散，但因篇幅冗长，造成最终的内积运算结果超过阈值，形成匹配成功的假象。

2.3 语义分析方法

基于关键词文本过滤的初步筛选减少了文本数量和文本语义分析的工作量，提高了过滤效率。事实上，只有对初步筛选的文本施行语义分析才有实际意义。近年来的研究表明，部分分析在大规模文本处理中是十分有效的，不仅降低了开销，从精确度和实际效果来看也是完全满足需要的。因此，在文本语义分析中，将文本划分为若干个分析单元，即设置窗口，根据关键词的位置，设置包含该关键词的一定句子长度的窗口，利用WordNet语义词典的汉语分析器，获取窗口中反映的各种语义关系，填充相应的局部语义框架，并生成文本的全局语义框架，与基于语义框架的用户模板相匹配，获得过滤模型的最终输出结果[5]。

2.3.1 窗口划分和局部语义框架的填充

收集用户模板中出现的关键词，并利用它将文本划分为不同的分析单元，即不同的局部语义框架。文本中含有多少单元就会有多少局部语义框架，而权重最大的局部语义框架称为全局语义框架。根据 WordNet 语义词典中语义关系与槽之间的映射规则来填充局部语义框架。

在 WordNet 语义词典中，语义关系用于在结构上描述句子节点间的关系及句子的符合关系，这些关系利用可区分不同关系的标号来表示，它是格关系的一种扩充。本文主要利用节点间的关系来描述特征项之间的关系。

由于表达手段的不同，文本中出现的关键词不一定出现在句子中，所反映的语义关系也不一定体现在同一个分析单元中，因此引入远程匹配机制，并规定相应的权重函数 $f_{\mathrm{dist}}(S)$，以便全面地反映窗口中体现的各类语义关系。为了体现不同槽在匹配中的地位，语义框架中各个槽的权重设置为 $f_{\mathrm{slot}}(S)$，则每个局部框架 F 的权重为

$$f_{\mathrm{local}}(F)=\frac{\sum_{i=1}^{l}f_{\mathrm{slot}}(S_i)\,f_{\mathrm{dist}}(S_i)}{\sum_{i=1}^{l}f_{\mathrm{slot}}(S_i)}$$

2.3.2 全局语义框架及文本与模板的匹配

设文本T的局部语义框架集合为$\{F_1, F_2,\cdots, F_M\}$，其全局语义框架为F_0，则取权重最大的局部语义框架F^*为F_0，其权重 $f_{\mathrm{globe}}(F_0)=f_{\mathrm{local}}(F^*)$。设用户基于语义框架的模板中所定义的阈值为 θ，当全局语义框架的最终权重 $f_{\mathrm{globe}}(F)>\theta$ 时，则认为该文本与用户模板匹配。最后，将符合条件的文本按照

权重的大小输出给相应的用户。

2.4 用户信息的反馈

随着文本过滤的进行，用户的需求可能会发生变化，为了适应这种动态的变化，系统需要有一个自适应学习的过程。利用用户的反馈信息及时对用户模板进行修改，以适应用户的变化，提高系统过滤性能。

对于输出的过滤文本，可以逐步收集用户的反馈意见，当相关文本达到一定数量时，应用潜在语义索引来改进文本过滤的效果，其主要步骤是：

(1)收集相关文本，形成初始的项/文本矩阵 $\boldsymbol{A}=(a_{ij})$。

$$a_{ij}=\frac{tf_{ij}\,\mathrm{lb}(N/n_j)}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2\left[\mathrm{lb}(N/n_k)\right]^2}} \qquad i=1,2,\cdots,m;\ j=1,2,\cdots,n$$

这是著名的$tf\times idf$公式的一个变形，主要考虑文本长度因素而进行归一化处理。其中，tf_{ik}表示关键词t_i在文本V_j内的频数；n_j表示文本集中包含关键词t_i的文本频数；N为文本集的数目。

(2)进行SVD分解，$\boldsymbol{A}\approx\boldsymbol{TSD}^{\mathrm{T}}$。

(3)新文本与用户模板的匹配。首先需要将文本和用户模板转化为潜在语义空间中的向量表示，然后计算其相似系数。此时，是在语义空间考察判断文本和模板之间的相似关系，而不是基于简单的词共现信息。

3 实验与结论

本文实验的测试文本集来自网站上搜索的百余篇文章，属于描述不同业务种类的网页文档。过滤方法分 2 种：(1)单纯采用关键词匹配的过滤方法；(2)采用本文提出的关键词匹配与语义框架匹配相结合的过滤方法。评估文本过滤模型的 2 个重要指标为：精确率(precision)和召回率(recall)。召回率是指已过滤出的相关文本占所有相关文本的比率，精确率是指在所有过滤出的文本中相关文本所占的比率。实验结果如图 2 所示。实验结果表明，本文的文本过滤模型综合了关键词匹配和语义框架匹配的优点，不仅适应了网络用户对信息需求实时性的要求，同时借助了语义框架匹配的特点，使得文本过滤的准确性得到了很大的提高。

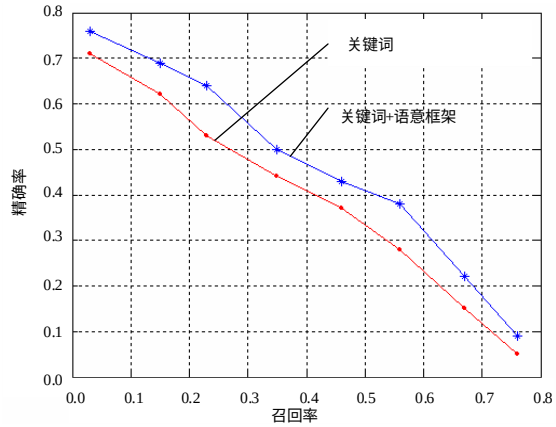

图 2　2 种文本过滤方法的性能比较

由于本文采用的语义框架匹配算法建立在关键词匹配算法的基础之上，因此其复杂度不会影响整个系统的过滤速度。目前应用广泛的网络安全技术——防火墙和隔离网闸对内容过滤的实时性和准确性都有很高的要求，本文的混合模式过滤模型可以很好地适用于这些技术。

参考文献

[1] Dumais S T. Using LSI for Information Filtering[C]//Proc. of the 3rd Text Retrieval Conference. [S. l.]: National Institute of Standards and Technology, 1995.

[2] 黄萱菁，夏迎炬，吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报, 2003, 14(3): 435-442.

[3] 何　静，刘海燕，张惠民. 基于文本的内容过滤算法的比较[J]. 计算机工程, 2002, 28(11): 9-11.

[4] Fillmore C J. The Case of Case[M]//Bach E, Harms R. Universals in Linguistic Theory. New York, USA: Holt, Rinehart and Winston, 1986.

[5] Laham D. Latent Semantic Analysis Approaches to Categorization[C]//Proceedings of the 19 the Annual Meeting of the Cognitive Science Society. New York, USA: Ablex Press, 1997.

(上接第 63 页)

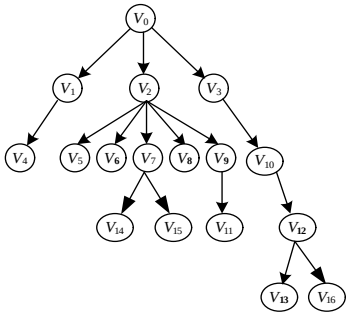

图 6　版本模型

5 结束语

版本管理是协同设计中一种用于跟踪和修改版本的方法。本文从数据存储的角度讨论了版本管理的方法过程，通过设置版本节点的存储阈值将差值存储与完整存储相结合，从而有效地平衡空间效率和时间效率，大大节省了存储空间，数据的安全性及版本存取速度得到了保证。

参考文献

[1] Roseman M A, Gero J S. Modeling Multiple Views in a Collaborative Environment[J]. Computer-aided Design, 1996, 28(3): 45-48.

[2] ISO 10303-1-1994(E) Industrial Automation Systems and Integration: Product Data Representation and Exchange——Part 1: Overview and Fundamental Principles[S]. 1994.

[3] 王云鹏，郭学旭. 计算机辅助协同设计中的数据管理研究[J]. 计算机辅助设计与图形学学报, 2003, 15(11): 21-25.

[4] 陈　虔. 大型应用软件的协同开发和管理机制研究——基于组件增量的版本控制模型 CDOM[D]. 北京：中国科学院高能物理所计算中心, 1998.

[5] Shreekanth M. Integrating the CAD Model with Dynamic Simulation: Simulation Data Exchange[J]. Concurrent Engineering Research and Application, 2002, 10(3): 239-250.

[6] Schewe K D, Thalheim B. Fundamental Concepts of Object Oriented Databases[J]. Acta Cybernetica Szeged, 1993, 11(4): 49-84.