基于概念的信息过滤技术探讨*

李荣陆　张永奎 牛伟霞
山西大学计算机科学系（太原　030006）
E-mail:lrl@sxu.edu.cn

摘要： 作为智能信息检索的一个分支，基于概念的信息过滤技术从语义级进行过滤文档和用户模型的匹配。本文对基于概念的信息过滤模型进行了描述，介绍了概念层的功能和构造方法，提出了一种基于概念的信息过滤系统的实现方法。最后，说明了当前基于概念的过滤系统存在的难点。

关键词： 信息过滤、用户模型、概念

A Survey of Conceptual Information Filtering

Li Ronglu　Zhang Yongkui　Niu Weixia
Dept. of Computer Science, Shanxi University, TaiYuan 030006
Email:lrl@sxu.edu.cn

ABSTRACT: As a branch of intelligent information retrieval, conceptual information filtering match the semantic content of documents with the semantic content of user profile. In this paper, the model of conceptual information filtering is described. The function and method of construction of the concept layer is introduced. And a method of realizing conceptual information filtering system is represented. Finally, the difficulties of conceptual information filtering are represented.

Keywords: information filtering, user profile, concept

1 引言

近年来信息检索和过滤的研究表明，要想使检索和过滤的性能得到显著的提高，必须使用一些技术以理解过滤文档的内容和用户的兴趣主题。基于这种思想，人们开发了一些智能检索系统，有的将专家系统与检索结合起来[1]，有的使用自然语言处理技术来进行检索[2,3]，还有基于概念和知识表示的检索系统，如 GRANT[4]和 RUBRIC[5]。由于一些

* 本课题得到山西省自然科学基金项目(991035)、山西省归国人员基金项目资助。
作者简介：张永奎，男，教授，博士导师，，主要研究领域为中文信息处理与人工智能。李荣陆，男，硕士研究生，研究领域为中文信息处理。牛伟霞，女，硕士研究生，研究领域为中文信息处理。

大规模语义词典的出现，如：WordNet 和 HowNet，使基于概念的检索和过滤系统实现起来较为容易，所以这方面的研究尤为引人注目。

基于关键字的机械式匹配，由于参与匹配的是字符的外在形式，而不是它们表达的概念，所以经常出现过滤不全、答非所问的结果。基于概念的过滤突破了机械式匹配局限于表面形式的缺陷，从词所表达的概念意义层次上来认识和处理过滤文档和用户的兴趣主题，在一定程度上表达了过滤文档和用户的兴趣主题语义信息，缩小了用户描述自己兴趣主题用词与文档索引词间的差距，从而提高了过滤系统的查全率和查准率。

2 概念过滤模型

我们把信息过滤系统 IF 形式化描述为[6]：

$$IF = (U, \overline{U}, D, \overline{D}, T, O, \rho)$$

其中，

U 表示用户兴趣模型（即用户模型 User Profile）实体的状态集，描述了它们在与系统交互时被创建、更新直至逐步完善的过程；$\overline{U}$ 表示用户兴趣模型描述符的状态集，描述了它们在与系统交互时，由于反馈作用的存在，其初始特征不断被更新的过程；D 表示由过滤系统新收集到的待过滤对象流集；$\overline{D}$ 表示由过滤系统新收集到的待过滤对象流的特征描述集；T 表示过滤系统中特征项的状态集，描述了过滤系统中由于新收集对象的加入，其特征项集不断被扩充的过程；O 表示过滤系统在发送待过滤对象和更新用户兴趣模型过程中生成的一系列输出结果集；ρ 表示相关度计算函数。

用户模型是使用用户模型描述符表示的用户兴趣主题。不同人常常会使用不同的描述符来表示相同的兴趣主题，而且同一个人在不同的时刻往往也会使用不同的描述符来表示自己的兴趣主题。文档是用文档描述符表示的潜在地符合用户兴趣的信息载体。同样，不同的人在表示相同的文档时，相同的人在不同时刻表示同一文档时，常常会使用不同的文档描述来表示文档。

在基于概念的过滤系统，用户模型和文档的匹配发生在概念层。在概念层，概念间通过带有权重的边连接起来，表示各个概念间的相关程度，用户模型描述符和文档描述符连接到自己所描述的概念层的概念上。概念层独立于用户模型描述符和文档描述符，这样即使用户模型描述符和用户模型描述符使用不同的表示形式，仍然可以进行用户模型和文档的匹配。同时，不同的用户模型描述符或文档描述符可以指向相同的概念，解决了过滤中多义和同义现象。

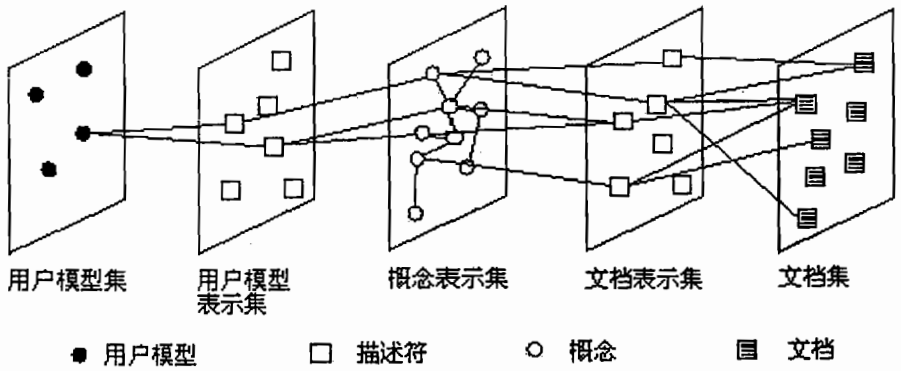

图 1　基于概念的过滤模型

根据概念层节点间的关系，可以使用语义网（Semantic Network）、联想网（Associative Network）、推理网（Inference Network）来表示概念层。

如果概念层节点间的连接代表的是语义关系，则概念层网络就可以看作是一个语义网，通常在要按照语义对节点间的连接上进行标注。而基于概念词典的系统，实质上是基于语义网的一种特例。

如果概念层节点间的连接代表的是普通意义上的关联，则概念层网络就可以看作是一个联想网。联想网节点间的连接不需要进行标注，只需加上权重，以表示连接的紧密度。

如果概念层节点间的连接代表的是逻辑蕴含（logic implication），则概念层网络就可以看作是一个推理网。它是由 Turtle 和 Croft[7,8]提出的，它的优点是具有很好的数学理论基础，可以将多种处理技术应用到它的上面。

3 概念模型

概念是人们对客观事物本质特征的概括，是一些相近实例的聚类，它必须能够表达这些实例的公共特征，一般我们通过字、词、词组等概念描述元素将其表达出来。

一般来说，概念包括以下三种信息。

(1)上层概念（概括性）；

(2)下层概念（专门性）；

(3)所有实例的公共特征或限制（选择性限制）。

其中，(1)，(2)是结构方面的信息；(3)是语义方面的信息。

3.1 概念层的功能

概念主要用来实现两个方面的功能：同义扩展和相关概念联想。

同义扩展用来解决人们在表达相同概念时，使用不同的词汇而造成不能满足关键词匹配的要求，如："计算机"、"微机"、"电脑"都可以表达相同的概念。相关概念联想用来解决概念间的相关性，因为概念不是独立存在的，它总是与其它概念间存在各种各样

的联系，人们希望过滤得到的不仅仅是文档，还希望能帮他产生一些新想法、提供一些建议、发现一些被遗忘的名称等等。这样，同义扩展可以提高过滤的查全率，相关概念联想可以加强过滤系统与用户的交互，使过滤系统在一定程度上可以表示出过滤文档和用户模型的语义信息，具有一定程度的智能。

3.2 概念层的构造

基于概念的信息过滤系统主要是使用概念层的知识去理解和细化文档和用户模型，所以其核心是概念层知识的构造。根据概念层的知识是否可以动态更新，概念层的构造分为两种方法。

(1)概念层的知识为静态知识。此方法认为领域知识的细节可以获得，概念层构造好后，不再进行修改，它往往使用于面向领域的信息过滤系统。对于这种方法人们常常使用层次结构来构造概念层，每一概念都可以细分为几个子概念，从最高层概念逐步细分，就形成了一个树状的层次结构。较低层的概念能自动继承较高层概念的全部特征。这样，我们就能把共享信息放在树状结构尽可能高的层次上，以减少数据冗余，提高存储效率。

(2)概念层的知识为动态知识。此方法认为领域知识无法完全获得，需要系统不断和用户进行交互学习，这种方法可以用于较宽的领域，概念层的构造可以使用人工神经网络等方法，而与用户交互学习的过程中可以使用各种机器学习的方法。

4 概念过滤系统实现

概念过滤系统的实现方法有很多种，我们使用由上往下逐步求精的策略来定义用户的兴趣主题概念。首先，将相应的用户请求表示为一个单一的词表示的概念，然后将此概念进行分解，通过一些逻辑上是“与”或“或”关系的下位词细化概念的描述，这些下位词或者是一个不同抽象层的概念，或者是索引词。上位词和下位词间具有一个权重值，代表下位词与上位词的相关程度。

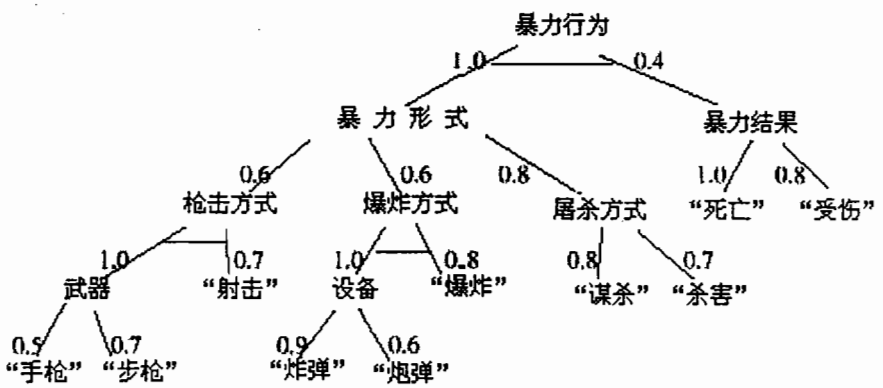

图 2　概念“暴力行为”的表示形式

图 2 显示了概念“暴力行为”的表示形式，叶子结点表示索引词，并且使用双引号括了起来，中间结点是一些其它概念，权重显示在连接边上。分支间使用一条线连接起来时，表示它们间是“与”关系；否则，表示是“或”关系。文档的相关度 RSV 通过一种由底向上的方法来计算。

例如文档中包含概念中的词手枪，射击，炸弹，杀害，死亡，那么这些索引词权重 1.0，其它文档中没有包括的词赋予权重 0。概念武器由两个下位词构成，手枪的权重为 1.0，步枪的权重为 0，它们间的关系是“或”关系，所以概念武器与概念暴力行为的相关度为 Max(1.0*0.5,0*0.7)=0.5，即所有下位词的权重与其连接权重的乘积中的最大值。概念枪击方式的权重可以使用同样的方法来计算，由于枪击方式的下位词间为与关系，所以应取最小值 Min(0.5*1.0,0.7*1.0)=0.5。最后，我们可以计算得到文档与概念暴力行为的相关度 RSV 为 0.3。

5 结束语

基于概念的信息过滤系统能够提供比基于关键字的机械式匹配更为智能化、知识化的服务，其根本基础在于拥有丰富的知识。因此概念层知识库的构造是基于概念的信息过滤系统的关键，现在仍有许多问题需要解决。

⑴增量化问题。由于知识的获取是一个动态的过程，它总是在不断更新的，所以必须保持知识库的知识可以不断更新。

⑵准确性问题。概念层知识的获取往往是使用统计的方法，或由专家描述获得的，这常常无法反映用户对知识的理解。

⑶颗粒度问题。在表示知识时，很难决定知识的详细程度，需要表示的程度越细，知识库越庞大。知识库太小的话，提供的语义信息非常有限，对文档和用户模型的理解和细化作用不大。

概念层知识库的构造已经成为了概念信息过滤的瓶颈，它的研究将是一个艰巨而费时的任务。

参考文献

[1] R.H. Thompson. The design and implementation of an intelligent interface for Information Retrieval. Technical report, Computer and Information Science Department, University of Massachusetts, 1989.

[2] Tomek Strzalkowski, Gees Strin, G. Bowden Wise, Jose Perez-Carballo, Pasi Tapanainen, Timo Jarvinen, Atro Voutillainen and Jussi Karigren. Natrual Language Information Retrieval:TREC-7 Report. Proceedings of TREC-7 conference. 1999.

[3] Tomek Strzalkowski, Jose Perez-Carballo, Jussi Karigren, Anette Hulth and Pasi Tapanainen&Timo Lathinen. Natrual Language Information Retrieval:TREC-8 Report. Proceedings of TREC-8 conference. 1999.

[4] P.R. Cohen and R. Kjeldsen. Information Retrieval by constrained spreading activation on Sematic Networks. Information Processing &Management, 23(4):249-254, 1987.

[5] R.M. Tong, L.A. Appelbaum, V.N. Askman, and J.F. Cunningham. Conceptual Information Retrieval using RUBRIC. In Proceedings of ACM SIGIR, New Orleans, Luisiana, June 1987.

[6] 张永奎、郭文宏、牛伟霞、李荣陆：网上中文信息过滤技术的研究，第一届中文信息处理发展国际研讨会，2001.4，上海。

[7] H. Turtle and W. B. Croft. Inference networks for document Retrieval. In Proceedings of ACM SIGIR, Brussels, Belgium, September 1990.

[8] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 9(3):187-222, July 1991.