一种基于本体的混合检索方法

杨学兵,孙 航
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:提出了一种基于本体的,综合改进的 spread activation 算法和语义分析的混合检索方法。通过改进的 spread activation 算法和本体实例之间语义关联强弱的分析,得到一组查询词的相似词集合,从而提高了查询关键字到本体概念映射的完整性与准确性。设计实现了相应的检索系统,实验表明,该系统可以有效地提高检索的查全率与查准率。

关键词:spread activation;语义关联;语义检索
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)01-0125-03

A Hybrid Retrieval Method Based on Ontology

YANG Xue-bing, SUN Hang
(School of Computer Science of Anhui University of Technology, Maanshan 243002, China)

Abstract: Bring forward a hybrid retrieval method which integrates the spread activation arithmetic and semantic analysis based on ontology. The semantic analysis between the modified spread activation arithmetic and ontological samples can get the similar words aggregate of the words you want to search for to enhance the integrity and accuracy of the ontology mapping of keywords. The design can realize the corresponding retrieval system, and the experiment also indicates that this system can improve recall ratio and precision ratio effectively.

Key words: spread activation; semantic relevance; semantic retrieval

0 引 言

随着网络所能提供的信息内容、网络结构、网络能力提供方法的不断发展,网络的规模有了爆炸性的增长,搜索引擎成了帮助人们寻找相关信息的重要工具。目前的搜索引擎普遍存在查全率和查准率不高的现象,任何一个简单的查询都可能返回数以万计的检索结果,而其中只有很少一部分与查询需求相关。

语义网[1]被认为是下一代的网络技术,它的核心是用元数据描述网络上的资源,使机器能理解网页的内容。语义网的技术已经广泛地应用到了搜索引擎。基于资源描述框架的问答系统(QuizRDF)[2]结合了传统的纯文本的搜索和 RDF 注释的资源查询、导航技术,能比较快速地收敛到用户查询的目标,但它不具备发掘概念间关系的能力。作者 Rocha C, Schwabe D[3] 以及 Guba R, McCool R[4]将用户输入映射到本体知识库,通过本体关系推导,发现与用户输入相关的概念。其中,作者 Rocha C, Schwabe D[3]提出的一种基于语义网的检索方法可以有效提高查询的查全率。但在关键字到本体库映射过程中,因为不能确定用户的意图,可能产生与用户查询关系不大的以及错误的关键字到本体实例的映射,导致查准率有了一定程度的下降。Anyanwu K, Sheth A.[5]提出了语义关联的概念,利用此概念,能有效提高搜索的准确性。梅翔,孟祥武[6]提出的一种基于语义关联的查询优化方法,就是利用语义关联的概念,将词法关系和语义分析相结合,在基于本体应用的基础上,产生用于传统搜索的关键字。与利用用户直接输入的查询关键字相比,在扩展性和精确性方面都有一定程度的提高,其中语义分析的方法提高了搜索的精确性。

结合语义关联的方法的优点,对 Rocha C, Schwabe D 提出的检索方法进行改进,提出了一种综合 spread activation[7]和语义分析的混合检索方法。

1 算 法

1.1 Weight Mapping 技术

本体及其本体实例中,有很大一部分信息是隐藏在它们之间的关系中,而不是明显表达的。在传统本体中只能指出两个概念实例之间的关系存在与否。而很多情况下,指出关系之间的权重也是很有必要的。一种经典的方法就是在本体实例之间的链接上关联数

收稿日期:2008-06-05
基金项目:安徽省自然科学基金重点资助项目(2004KJ053ZD)
作者简介:杨学兵(1967-),男,安徽巢湖人,教授,研究方向为数据挖掘。

字值。Weight Mapping 技术就是在表示关系属性的连接上赋予一定的权值，用来代表不同的关系属性具有不同的权重。提议两种赋权重的方法：Cluster measure 以及 Specificity measure。

Cluster measure 用来表示一个关系中的两个概念实例的相似度，相应公式如下：

$$W(C_j,C_k)=\frac{\sum_{i=1}^{n}n_{ijk}}{\sum_{i=1}^{n}n_{ij}},C_j,C_k\text{ 为对应的概念实例。}$$

n_{ij} 表示 C_i 与 C_j 是有关的(如果两者之间有关，n_{ij} 为 1，否则，为 0)。n_{ijk} 表示 C_j，C_k 都和 C_i 有关(如果 C_j，C_k 都和 C_i 有关，n_{ijk} 为 1，否则，为 0)

Specificity measure 用来表示关系的特异性，相应公式如下：

$$W(C_j,C_k)=\frac{1}{\sqrt{n_k}},n_k\text{ 的值为给定的关系类型的实例的数量，该关系类型以结点 }k\text{ 为终结点。}$$

1.2 本体语义关联

目前在信息检索领域用概念关联度来衡量概念间的联系，从自然语言的角度来讲主要考虑语义相关度和语义相似度两方面因素。

语义相似度是指概念在意义上的相符合程度，在语义树中通过概念的语义距离计算语义相似度，概念的语义距离与语义相似度成反比。设有概念 C_1 和 C_2，其语义相似度计算方法为：

$$\mathrm{sim}(C_1,C_2)=1-t\sqrt{\frac{1}{2}\alpha\mathrm{Dist}(C_1,C_2)}$$

其中 $\alpha=\dfrac{\mathrm{Dep}(C_2)}{\mathrm{Dep}(C_1)+\mathrm{Dep}(C_2)}$，$\mathrm{Dep}(C)=\sum_{i=1}^{n}1$，

$$\mathrm{Dis}(C_1,C_2)=\sum_{i=1}^{n}\frac{1}{\mathrm{Wid}(C_i)}\frac{1}{2^{\mathrm{Dep}(C_i)}}$$

语义相关度是指概念在语义上的关联程度，如“医生”和“病人”则属于关联关系。语义关联关系在语义树中表示为两个概念间是否存在路径，路径的长度表明概念间关联程度。设有概念 C_1 和 C_2，语义相关度计算法为：

$$\mathrm{Relativity}(C_1,C_2)=\frac{r}{\mathrm{length}(C_1,C_2)+r}$$

其中 r 为调节参数。

语义关联度受语义相似度和语义相关度的共同影响，概念 C_1 与 C_2 间的关联度可表示为：

$$\mathrm{Crelevancy}(C_1,C_2)=\alpha\mathrm{Sim}(C_1,C_2)+\beta\mathrm{Relativity}(C_1,C_2)\quad\text{其中 }\alpha+\beta=1$$

1.3 混合检索算法

Spread Activation 算法[8]是最常用的语义网络的处理框架之一，被成功地应用于很多领域，特别是在信息抽取(Information Retrieval)方面的应用[5,9]。从它在人工智能领域用于作为对语义网络和本体的处理框架开始，就把它视为在语义网领域的一个自然和有趣的知识处理算法。这个算法的基本原理类似于一个概念探索器，善于进行相似性查找。

将 Weight Mapping 技术与传统的 Spread Activation 算法相结合，就使 Spread Activation 算法得到改进。通过这种改进的 Spread Activation 算法扩展用户的查询，可以很大程度上提高查询的扩展性。但带来扩展性的同时，也会出现大量的无用数据，使用户的查询失去意义。这里利用语义关联的优点，与此算法相结合，就可以在提高扩展性的同时，提高查询的精确度。具体的计算方法如下：

1)通过查询关键词映射到相应的本体实例，将这些本体实例作为初始化结点集合，$N=\{n_1,n_2,\cdots,n_k\}$，n_i 为用户输入的第 i 个关键词，k 为关键词个数。每个初始化结点被赋予一个激发值，作为算法的输入参数。可以根据这些初始化结点的重要性，设置不同权重的激发值。按照这些激发值的大小，以非递增的方式将这些初始化结点放到一个优先队列中。具有最高激发值的结点就会最先被拿出进行处理。如果现在拿出的节点通过所有的约束，它就会将其激发值传给其邻接点。假设初始结点是 I，终结点为 j，向邻接点的传播通过以下公式作用，I 代表输入，O 代表输出：

$$I_j(t+1)=O_i(t)*w_{ij}*(1-\alpha)$$

函数 $O_i(t)$ 代表结点 i 的输出，这里 $O_i(t)$ 选择线性函数(一个结点的输出值等于其输入值)。w_{ij} 代表通过 Weight Mapping 技术得到的值。每次传播，都会将激发值的损失考虑进去。$1-\alpha$ 代表每次传播激发值损失的比例。将激发值传给邻接点，就说明邻接点也被激发，如果该邻接点不在优先队列中，就会被加进去，然后对优先队列按照激发值的大小重新排列。被从优先队列拿出并处理过的结点会放到一个结果队列中。这个处理过程会一直重复下去，直到满足了定义好的一种状态，或者在优先队列中已经没有待处理的结点。

2) 得到每个关键词 $n_i(1\leqslant i\leqslant k)$ 在本体中的相近词集合 $S'_i=\{s_{i1},s_{i2},\cdots,s_{im}\}$，$m$ 为 n_i 的相近词个数，共得到 k 个相近词集合。

3) 计算 s_{ij} 与关键词 n_i 之间的关联程度 $V_{ij}=\alpha\mathrm{Action}(j)+\beta\mathrm{Crelevancy}(n_i,s_{ij})$，$\mathrm{Action}(j)$ 为结点 j 的激发值，s_{ij} 为集合 S'_i 中的一个结点，$\alpha+\beta=1$。当满足 $V_{ij}\geqslant V_{\min}$ 时，将 s_{ij} 放入最终的相似词集合 S_i 中；否则，认为 s_{ij} 与关键词 n_i 的关联程度达不到要求，不放入集合 S_i 中。经此过滤操作后，得到最终的相似词集

合 S_i, S_i 中的元素以关联程度值的大小为依据,非递减排列。

用户可以根据需求对 α, β, V_{min} 的取值进行配置。

2 查询优化系统实现

实现的查询优化系统如图 1 所示。

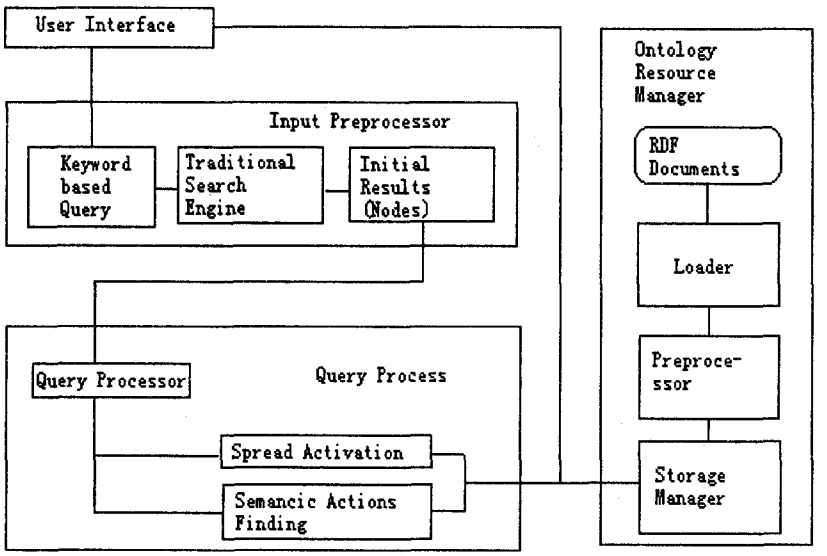

图 1　查询优化系统结构

系统分为 User Interface, Input Preprocessor, Query Process, Ontology Resource Manager4 个部分。

(1)User Interface:负责人机接口,接受用户查询请求。

(2)Input Preprocessor:对输入的关键词进行预处理,使用传统的搜索引擎的方法在指定的本体中匹配到相应的本体实例。

(3)Ontology Resource Manager:维护和管理语义优化依赖的本体知识。本模块把本体知识库中以资源描述框架(RDF)格式描述的本体实例及其之间的关系转化为内部的数据结构。

(4)Query Processor:对匹配到的本体实例进行扩展,并进行本体间语义关联的发现与评价,将结果返回给用户界面。

3 性能测试

利用 Department of Informatics at PUC－Rio 的网站进行测试。在此网站中,可以得到主要测试领域的信息:professors, projects, students, labs and publications。相关知识库包含 2630 个本体实例结点和 6554 个关系实例。在此网站中,要用于进行测试本体的一小部分如图 2 所示。

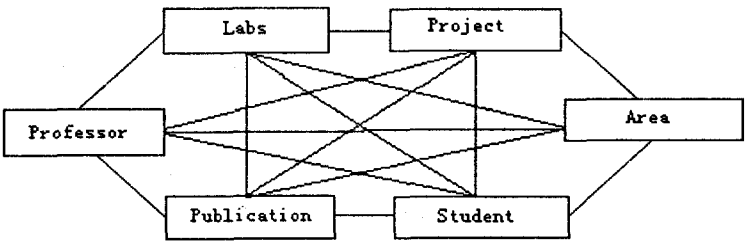

图 2　部分本体知识库实例

此网站包括 10 个研究领域,如软件工程、人工智能、数据库等等。每个领域还有不同数量的分支领域。

假设一名学生想要考取人工智能中的一个研究方向数据挖掘,他首先要了解在这方面发表文章比较多的教授有哪些。但是由于数据挖掘只是人工智能的一个子方向,在此系的网站中,没有作为正式的领域列出。相应地,该学生不能直接得到有关哪些教授在此领域发表的文章比较多等等信息。如果只是使用传统的搜索方式,他只能在该网站中输入关键词"数据挖掘",然后在结果结点中浏览,分析所有的出版物,找出哪些导师在这个领域的表现是比较突出的。这是很难找的,因为出版物可能有成百上千份。

利用文中提出的方法进行检索,输入关键字"数据挖掘",就会在输出结果中发现,在"数据挖掘"领域发表过很多文章的教授实例。这种混合的检索方法是通过遍历本体实例图做出推论,得出教授和"数据挖掘"领域相似度很高。同时传统的 Spread activation 算法会得出一定数量的无用数据,影响查询结果。而该混合检索方法,很大程度上降低了无用数据出现的数量,保证了高质量的查询结果。

以查询进行"数据挖掘"领域研究的教授的资料为例,验证文中提出方法的有效性。结果如图 3 所示。

算法	前 10 篇中符合条件的文档数	前 10 篇的查准率	前 100 篇符合条件的文档数	前 100 篇的查准率
传统 Spread Activation	7.56	75.6%	68.4	68.4%
混合检索算法	7.93	79.3%	73.4	73.4.%

图 3　查询结果情况

从实验结果看,文中算法的准确率在 100 篇文档中有所降低,其主要原因可能是实验选取的检索源数量相对较小,以及对 α, β, V_{min} 的设置没有达到最佳,影响了查准率。以后将进行更大数量的测试训练,在此基础上进行修改,争取进一步提高查准率。

(下转第 130 页)

3.2.4 注册中心

一个 ebXML 注册中心存储着各种各样的业务处理过程和协作草案,提倡共享和重复使用共同的组件,以节约执行的时间和增进互操作。

一个在服务器上的注册中心能够被访问。用户通过登记 CPP,发布电子商务信息,也可在注册中心收索合适的贸易伙伴[7]。

3.2.5 消息服务

ebXML 采用基于 SOAP 协议的安全的标准方式传送消息,并采用 XML 数字签名。ebXML 消息服务有相应的出错处理、同步应答机制。它通过使用确认、重试和双重检测以及排除机制,使接收方有且仅有一次地接收消息。它的可靠性通过接收方回复带有确认消息的方式实现。

4 结束语

ebXML 结合 EDI、XML 和 Web 服务的特点,构建了新一代的电子商务标准。其在电子病历等医疗方面上的应用,国内外都有许多成功的例子。许多企业、公司和政府等部门都积极投入到采用 ebXML 构建新型电子商务平台的队伍中。其实,不论是电子商务,还是电子政务的数据交换,采用 ebXML 不需考虑不同对象的数据差异性,方便进行数据交换,并保证信息的安全。要实现 ebXML 的平台,关键是实现文中提到的其几个组成部分。这些都是在开发中需考虑的问题。此外,在安全上,结合了 ebXML 的 XML 数字签名在电子商务中得到广泛的应用。

参考文献:

[1] Hill N C, Ferguson D M. Electronic Data Interchange: A Definition and Perspective[J]. The Journal of Electronic Data Interchange, 1989, 1(1): 5 - 12.

[2] ebXML Technical Architecture Project Team. ebXML Technical Architecture Specification v1. 0. 4 [EB/OL]. 2001. http://www.ebxml.org/specs/ebTA.pdf.

[3] Dogac A. Exploiting Semantic of Web Services through ebXML Registries [EB/OL]. eChallenges 2003. 2003 - 10. Bologna, Italy. http://www.srdc.metu.edu.tr/webpage/publications/2003/tutorial.pdf.

[4] Hofreiter B, Huemer C, Kim Ja - Hee. Choreography of ebXML business collaborations [J]. Inf. Syst. E - Business Management, 2006, 4(3): 221 - 243.

[5] ebXML Requirements Team. ebXML Requirements Specification Version 1.06[EB/OL]. 2001. http://www.ebxml.org/specs/ebREQ.pdf.

[6] Business Process Project Team. ebXML Business Process Specification Schema Version 1. 01 [EB/OL]. 2001. http://www.ebxml.org/specs/ebBPSS.pdf.

[7] ebXML Registry Project Team. ebXML Registry Services Specification v1. 0 [EB/OL]. 2001. http://www.ebxml.org/specs/ebRS.pdf.

(上接第 127 页)

4 结束语

文中提出了一种基于本体的综合 Spread Activation 算法和语义分析的检索方法。实验表明,该扩展算法,能有效扩展查询关键字的查询结果,达到了预期的效果,基本满足了实际应用需要。今后将在现有工作的基础上,对基于领域的语义关联计算方法等问题做进一步的研究。

参考文献:

[1] Berners - Lee T, Hendler J, Lassila O. The semantic Web[J]. Scientific American, 2001, 284(5): 34 - 43.

[2] Davies J, Weeks R, Krohn U. QuizRDF: search technology for the semantic Web[C]//2004 Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04). Washington: IEEE Computer Society, 2004: 112 - 119.

[3] Rocha C, Schwabe D. A hybrid approach for searching in the semantic Web [C]//Proceedings of the WWW 2004. New York: ACM Press, 2004: 374 - 383.

[4] Guba R, McCool R. Semantic search[C]//Proceeding of the WWW 2003. New York: ACM Press, 2003: 700 - 709.

[5] Anyanwu K, Sheth A. ρ - queries: enabling querying for semantic associations on the semantic Web[C]//Proceeding of the WWW 2003. New York: ACM Press, 2003: 690 - 699.

[6] 梅 翔,孟祥武. 一种基于语义关联的查询优化方法[J]. 北京邮电大学学报, 2006, 29(6): 107 - 110.

[7] Crestani F. Application of Spreading Activation Techniques in Information Retrieval[J]. Artificial Intelligence Review, 1997, 11(6): 453 - 482.

[8] O'Hara K, Alani H, Shadbolt N. Identifying Communities of Practices: Analyzing Ontologies as Networks to Support Community Recognition[C]//IFIP - WCC 2002. Montreal: Kluwer, 2002.

[9] Cohen P, Kjeldsen R. Information Retrieval by Constrained Spreading Activation on Semantic Networks[J]. Information Processing and Management, 1987, 23(4): 255 - 268.