

语义检索关键技术研究

胡 哲, 郑 诚, 王艳玲

(安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘 要:传统的基于关键字的信息检索,由于忽视了关键词本身所含的语义信息,故只能得到较低的查全率和查准率。而源于知识工程和人工智能领域的本体理论和技术,能够很好地处理自然语言理解问题和具有基于语义的推理机制,因此成为改进传统信息检索方式的良好工具。与传统的检索技术相比,它能提高检索的精度和覆盖率,减少了不相关的返回结果。文中将对语义检索系统中所涉及到的语义检索预处理及查询语义扩展等关键技术进行分析研究,为语义检索系统的顺利实施奠定了良好的理论与实践基础。

关键词:本体;语义检索;查询语义扩展

中图分类号:TP301.2

文献标识码:A

文章编号:1673-629X(2008)10-0075-04

Research of Key Technologies of Semantic Retrieval

HU Zhe, ZHENG Cheng, WANG Yan-ling

(School of Computer Science and Tech., Anhui Univ., Hefei 230039, China)

Abstract: Because of ignoring the semantic information inside the keywords, the traditional searching engine based on the key words has lower recall and precision. However, the theory and technology of ontology, which develops from fields of knowledge engineering and artificial intelligence, have the ability to process and understand natural language, as well as dealing with semantic problems. Therefore, ontology becomes a good tool in advancing traditional information retrieval. Compared with the traditional search, the semantic search results in improvement in the precision and the recall of search application. And it also decreases the irrelevant searching return. Here, aim at two key techniques—semantic pretreatment and semantic query expansion of semantic retrieval system which establishes both the theoretical and the practical basis for implementation of semantic search system.

Key words: ontology; semantic retrieval; semantic query expansion

0 引言

目前,大多数的信息检索方法主要是基于关键字的查找,这些传统的检索方法在信息检索的发展过程中占有非常重要的地位,但是它们的缺点却也非常明显。用户需要输入关于查找的精确关键字。关键字稍有偏差,检索系统就无法确定用户的真正需要,因而无法提供正确的结果。检索过程只是在进行形式上的匹配,即使用户输入了精确的关键字,计算机还是不能进行语义层次的检索,检索结果的质量并不能令人满意。

为了解决这些问题,研究者尝试从语义的角度进行考虑,提出了各种新的方法和技术,也取得了很多成

果。通常的研究主要从自然语言处理、基于概念的方法以及基于本体的思路等三个方面来实现语义在信息检索中的集成和应用。文中将从本体的思路入手,探讨建立语义检索系统中的关键技术。

1 语义检索技术的研究现状

在语义检索领域中,通常的研究主要从自然语言处理、基于概念的方法以及基于本体的思路等三个方面来实现语义在信息检索中的集成和应用。自然语言处理(NLP)技术试图通过将某个查询的语义信息与文档的语义信息进行匹配来提高查询的性能^[1]。Hsinchun Chen首先提出基于概念的文本自动分类与语义检索,采用机器学习的方法实现了大量文本自动分类、标注与检索。随后概念空间^[2]描述了概念及其关系的联系,根据概念之间的相互联系,形成蕴含语义的关系网。在关系网中,可以实现同义词扩展检索、语义蕴含扩展以及语义相关扩展等。最早在1994年 Voorhees就曾提出基于本体的查询扩展^[3],使用了本

收稿日期:2008-01-22

基金项目:安徽省自然科学基金资助项目(050420204);安徽省高校自然科学基金项目(2006kj055B)

作者简介:胡 哲(1984-),女,安徽合肥人,硕士研究生,研究方向为语义检索与数据挖掘;郑 诚,副教授,硕士研究生导师,研究方向为数据挖掘与语义 Web。

体中的概念进行查询扩展,并得出最有效的方式:利用本体中的同义词和特定的子类关系进行扩展。此后基于本体的查询扩展研究侧重于两个方面:基于结构化的方法和基于注释的方法。

语义检索主要是基于概念匹配的检索方法,把传统方法中从用户查询和文档抽取出来的关键词替换为含有语义的概念,以此把关键字/词级的检索提升到概念级的检索。部分语义检索的研究也考虑到了概念和概念之间的关系,这种方式对早期的检索效果有较大提升,具有相当的参考价值和实践意义。

然而,这些方法的侧重点要么是针对文档中出现的语义概念,要么是对用户查询所涉及的本体概念,而没有充分利用到本体中的属性和其它关系,仅仅对概念的出现次数和频率进行考虑,从实质上忽略了文档和用户查询中蕴含的其它有价值的信息。应当说这样的本体利用是粗糙的。换言之,如果能够把这些出现在文档和用户查询中的概念、关系以及属性等信息综合考虑,并使它们的价值在检索过程中得到体现,就能够更进一步把语义检索的作用发挥出来,这正是笔者研究的目的所在。

2 关键技术研究

关键技术研究主要集中在两个方面^[4]。首先,对用户的输入进行语义检索项预处理,转化为本体中的概念、属性(即关系)和实例,即获得待扩展的核心概念集合。然后,根据本体中的语义信息进行查询语义扩展,获得加权的扩展概念集。

在此,假设用户的输入为本体中已经定义的概念、属性和实例。研究者大多考虑了概念、实例的扩展而忽略了属性,但笔者则认为,这种忽略是有隐患的。用户输入的术语都是用户意向的载体,属性当然也是不应被简单地忽略。而且,在本体中,正是属性将概念之间的语义联系很好地体现出来,抛弃属性就等于抛弃了本体中概念之间的重要语义联系,实是不智之举。

2.1 语义检索项预处理

2.1.1 用户输入预处理

基于对用户输入假设,将用户输入术语集合 User 转化为本体中三个术语集合 O_s 、 O_p 、 O_o 。其中 O_s 为 User 中可以和本体主语(Subject)相对应的部分, O_p 为 User 中可以和本体属性(Property)相对应的部分, O_o 为 User 中可以和本体宾语(Object)相对应的部分。

算法 1:用户输入预处理

输入:用户输入检索词集合 User

输出:三个本体术语集合 O_s 、 O_p 、 O_o

Begin

- 1) 如果 User 不为空,依次取 User 中的元素 u ;
- 2) 如果 u 是主语,将 u 存入集合 O_s ;
- 3) 如果 u 是宾语,将 u 存入集合 O_o ;
- 4) 如果 u 是属性,将 u 存入集合 O_p ;
- 5) $User = User - u$, 如果 User 为空,结束,否则转

1;

End.

2.1.2 语义连接

笔者利用本体中的语义知识,挖掘用户输入检索词之间的语义联系,形成六种有意义的语义连接: SPO 、 SO 、 SP 、 PO 、 S 、 O 。

算法 2:语义连接算法

输入: O_s 、 O_p 、 O_o

输出: SPO 、 SP 、 SO 、 PO 、 S 、 O

Begin

- 1) 如果 O_p 不为空,依次取 O_p 中元素 p ;否则转 15);
- 2) 如果 O_s 不为空,依次取 O_s 中元素 s ;否则转 10);
- 3) 如果 $\{s, p, * \}$ 在本体中转 4),否则转 9);
- 4) 如果 O_o 不为空,依次取 O_o 中元素 o ;否则转 8);
- 5) 如果 $\{s, p, o\}$ 在本体中,将 $\{s, p, o\}$ 写入集合 SPO ;
- 6) 如果 O_o 中还有下一个元素,转 4);否则转 7);
- 7) 如果 O_o 集中有元素使得 $\{s, p, o\}$ 属于本体转 9),否则转 8);
- 8) 将 $\{s, p, * \}$ 写入集合 SP ;
- 9) 如果 O_s 中还有下一个元素转 2),否则转 11);
- 10) 如果 O_o 不为空,依次取 O_o 中的元素 o ,否则转 14);
- 11) $\{*, p, o\}$ 是否在本体中,是则转 12),否则转 13);
- 12) 将本体中所有的符合属性为 p ,宾语为 o 的三元组 $\{s, p, o\}$ 写入集合 PO ;
- 13) 如果 O_o 中还有下一个元素,转 10),否则转 14);
- 14) 如果 O_p 中还有下一个元素,转 1),否则转 15);
- 15) 如果 O_s 不为空,依次取元素 s ;否则转 23);
- 16) 如果 O_o 不为空,依次取元素 o ,否则转 24);
- 17) 检查 $\{s, *, o\}$ 是否在本体中,若在,加入集合 SO ;
- 18) 如果 O_o 中还有下一个元素转 16),否则转 19);

19) 如果 O_o 集合中存在元素使得 $\{s, *, o\}$ 属于本体转 22), 否则转 20);

20) 如果 s 没有任何一个元素 p , 可以与之组合转 21), 否则转 22);

21) 将 $\{s, *, * \}$ 写入集合 S ;

22) 如果 O_s 中还有下一个元素转 15), 否则转 23);

23) 如果 O_o 不空, 检查集合 O_o , 将所有没有 s , 且没有 p 可以成功组合的元素 o , 以 $\{*, *, o\}$ 的形式写入到集合 O 中; 结束;

24) 如果 O_s 不空, 将无 o 无 p 组合的 s , 以 $\{s, *, * \}$ 的形式写入集合 S ;

End.

2.1.3 核心概念提取

笔者在此给出这六种组合的处理方法:

1) $\{s, p, o\}$: 如果 o 是常量, 选择 s 为核心概念, 如果 o 为资源, 选择 o 为核心概念。

2) $\{s, *, o\}$: 如果 o 为常量, 选择 s 为核心概念, 如果 o 为资源, 选择 o 为核心概念。

3) $\{s, *, * \}$: 选择 s 为核心概念。

4) $\{*, *, o\}$: 如果 o 为资源, 选择 o 为核心概念, 如果 o 不是, 查找本体, 找出 o 的类型 s , 加入核心概念集。

5) $\{s, p, * \}$: 选择 s 为核心概念。

6) $\{*, p, o\}$: 如果 o 为资源, 选择 o 为核心概念。

此外, 一个常见问题, 如何处理这样一类输入: 类 (+ 属性) + 属性值, 扩展模块需要找出用户需要的实例。

(1) 类 + 属性 + 属性值: 在语义连接后, $\{*, \text{属性}, \text{属性值}\}$ 应属于集合 PO , 很显然, 这类三元组语句的主语不能被丢弃。应该把它们加入到核心概念集合中去。特别的, 如果找到的主语正是输入类的一个实例, 这个实例应赋予重要的权值。

(2) 类 + 属性值: 语义连接后, $\{s, *, * \}$ 应属于集合 S , $\{*, *, o\}$ 应属于集合 O 。显然, 以上 6 种处理, 不能找出用户需要的实例。首先, 要找出 s 所有的实例集合 $S1$ 。现在, $S1$ 的元素与 o 之间的关系可以在一个三元组语句的距离内找到。如果 $s1$ 是 $S1$ 的元素, 而 $\{s1, *, o\}$ 在本体中, 那么这个 $s1$ 也应该纳入核心概念集合并赋予重要权重。

这样的问题应该还有更多, 也应该不断地细化和深化, 而不应总是粗糙地利用本体。

2.2 查询语义扩展

为了对本体概念相关内容进行扩展, 首先考虑利用本体中的层次结构关系来进行处理。查询中常用到

的本体中关系有: 同义词关系、父子关系、子树节点、兄弟节点、兄弟子树节点。通常情况下, 研究者考虑的是利用父子关系的节点以及子树节点来对查询概念进行扩展, 但忽略了属性。进一步的, 笔者同时考虑对这两方面进行扩展。

首先, 为本体中的概念建立一个连通图^[5], 在图中每一个边都有相应的权值, 作为概念相互扩展的语义参考。

算法 3 本体概念连通图构造算法

输入: 领域本体, 训练文档集

输出: 本体概念连通图

Begin

1) 初始化本体概念连通图;

2) 取相应本体领域的训练文档集, 统计每一篇文档中出现的概念数;

3) 依次取文档 D_t ;

4) 对 D_t 中出现的任两个概念 C_i 和 C_j , 取其出现频率小者, 作为 C_i 和 C_j 同时出现在文档 D_t 中的次数 f ;

5) 如果 C_i 和 C_j 在本体概念图中连通, 则把次数 f 累加;

6) 如果 C_i 和 C_j 不连通, 则连通 C_i 和 C_j , 并为其赋值 f ;

7) 如果还有没有计算的文档转 3), 否则转 8);

8) 对图中所有边的次数取最大值作为分母进行归一化, 得到连通图权重 $\omega_{i,j}$;

9) 将本体中的语义信息加载至概念连通图中, 遍历本体, 按照概念间不同的属性联系, 向连通图中对应的边追加不同的权值;

10) 对图中所有的边取权值最大值作为分母进行归一化, 得到连通图权重。

End.

算法的关键在于第 9 步, 本体语义信息的添加。本体语义信息的添加有多种方式, 在实验中, 笔者选择为领域本体的不同属性定制不同的权值。在实际应用中, 这些权值的设定应该有领域专家的参与, 经过大量的实验, 确定最佳的权值。对核心概念进行查询语义扩展的算法如下:

算法 4 查询语义扩展算法

输入: 本体概念连通图, 核心概念

输出: 概念扩展结果

Begin

1) 加载本体概念连通图;

2) 如果有需要扩展的概念 C , 则执行步骤 3), 否则结束;

3) 对概念 C 的相邻节点按权值从高至低排列, 取前 k 个作为 C 的扩展, k 为预先设定的参数, 如果在扩展概念集中, 该概念已经存在, 累加权值, 否则将该概念及权值加入扩展概念集; 转步骤 2)。

End.

3 实验分析

实验条件: Protégé 本体建立工具, Java 编程语言, Eclipse 平台, Jena API, Lucene, JSP, 100 篇期刊论文。

实验内容: 利用 Protégé 建立数据结构的领域本体, 利用 Jena API 和 Lucene 实现第二部分的所有算法, 建立一个相对完整的查询语义扩展模块。

实验模块结构及核心模块展开分别如图 1、图 2 所示。

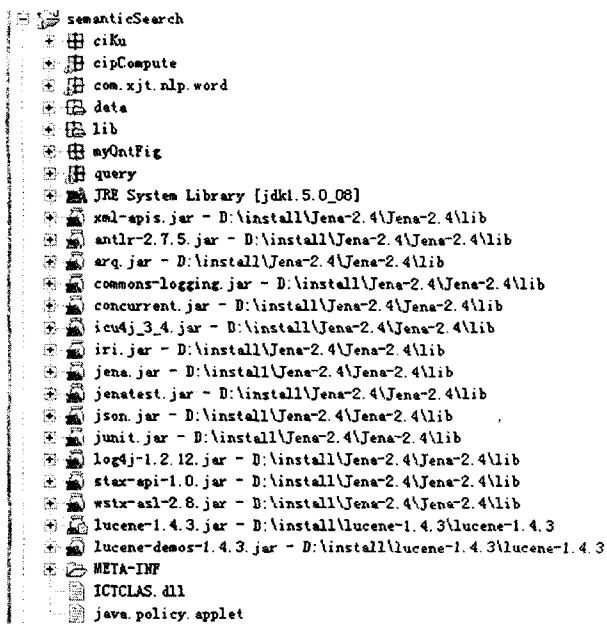


图 1 模块结构图

其中, searchColeConcepts 实现了算法 1, ConceptPreTreatment 实现了算法 2, cipCompute 和 myOntFig 实现了算法 3, ExpConcepts 实现了算法 4。

笔者列出两种典型输入:

实验结果:

用户输入: 数组 二维数组

部分结果:

概念	二维数组	数组的数组	数组	矩阵	动态矩阵	字符串数组
权重	1.0	1.0	1.0	0.8	0.42	0.42

相关的部分本体三元组:

数组的数组	存储方式	顺序存储;
数组的数组	sameAs	二维数组;
数组的数组	type	数组
二维数组	存储方式	顺序存储
二维数组	sameAs	数组的数组

二维数组	type	数组
矩阵	subClassOf	数组

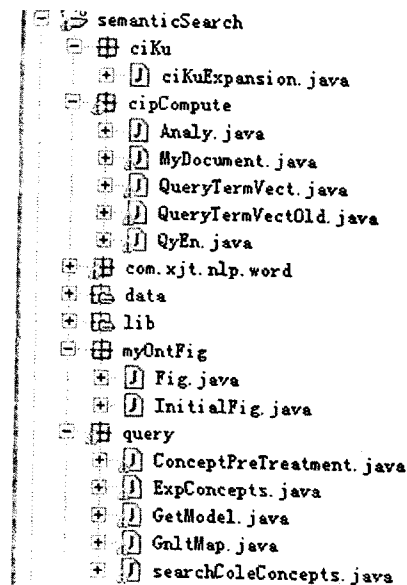


图 2 核心模块展开

这个输入着重考察模块对本体中的语义信息处理能力。在本体中, 数组的数组和二维数组是相等的实例, 数组因为父子关系获得了高的权值, 字符串数组动态矩阵因为语义关联较小所以获得较小的权重。

用户输入: 线性表 顺序存储 存储方式(可选)

部分结果:

概念	线性表	顺序表	静态链表	顺序存储
权重	1.0	1.0	1.0	0.8

相关的部分本体三元组:

顺序表	存储方式	顺序存储
顺序表	type	线性表
静态链表	存储方式	顺序存储
静态链表	type	线性表

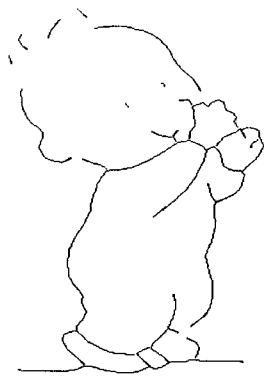
这个输入意在考察模块处理语义连接的功能。显然, 模块已经成功地找出顺序表、静态链表概念并赋予重要权重。

由实验结果可知, 该查询语义扩展模块可以更细致地利用本体中的语义背景知识, 扩展的结果更为全面科学, 对查询语义扩展后的概念进行检索, 会提高检索的查准率和查全率。

4 结束语

介绍了语义检索技术的研究现状, 尝试从语义的角度进行考虑, 从本体的思路入手, 研究语义检索系统中的关键技术, 建立了查询语义扩展模块。实验证明, 该查询语义扩展模块可以更细致地利用本体中的语义背景知识进行语义层面上的扩展, 从而达到提高检索查准率和查全率的目的。

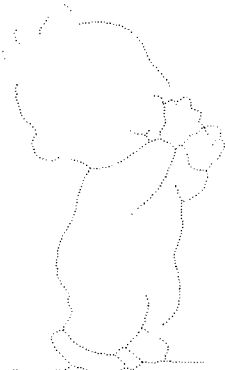
(下转第 81 页)



(a) 文中算法从图像中提取出的线条画



(b) 间隔12个点采样的结果



(c) 间隔3个点采样的结果

图 5 采样稀密不同时获得的笔划控制点

采用的公式如下:

$$P_i^{(n+1)} = P_i^{(n)} \pm F(k_i^{(n)}) * N_i^{(n)}$$

其中 $P_i^{(n)}$ 为特征点 i 第 n 次移动的结果, k_i 为点 i 处角度大小, 代表该点的曲率, N_i 为角平分线方向, 代表法线方向。通过改变 N_i 的方向, 对曲线进行平滑或夸张, 其夸张幅度也可以通过调整 $F(k_i^{(n)})$ 的大小控制。夸张风格的线条笔划效果如图 6 所示。

3 结束语

文中主要完成的工作是从卡通画中提取出具有表现力的矢量线条。从实验结果看, 从原图像抽象产生

的线条画, 保留了原图像的主要特征, 是原图像内容的简洁表示。但文中的方法还不能适用于所有的图像, 其范围有一定的限制。因而, 在此基础上可以依据一幅具有某种风格的线条画作为样本, 把图像转换为具有该样本风格的线条画^[4,5]。

探寻线条画风格转换和风格定制的方法, 是以后研究工作的重点。



图 6 具有夸张风格的线条笔划

参考文献:

- [1] DeCarlo D, Santella A. Stylization and abstraction of photographs [J]. ACM Transactions on Graphics, 2002, 21(3): 769-776.
- [2] Fang Wen, Qing Luan, Lin Liang, et al. Color sketch generation [J]. Non-Photorealistic Animation and Rendering, 2006, 5(7): 47-54.
- [3] DeCarlo D, Finkelstein A, Rusinkiewicz S, et al. Suggestive contours for conveying shape [J]. ACM Transactions on Graphics, 2003, 22(3): 848-855.
- [4] 孙玉红, 屠长河, 孟祥旭. 基于形状演化的线条画风格转换与变形 [J]. 计算机辅助设计与图形学学报, 2006, 18(2): 208-211.
- [5] 屠长河, 孙玉红, 孟祥旭. 基于样本的线条画风格转换与定制方法的研究 [J]. 计算机学报, 2005, 28(6): 965-971.

(上接第 78 页)

参考文献:

- [1] Strzalsowski T. Natural language processing in large-scale text retrieval tasks [C] // The First Text Retrieval Conference (TREC-1). Gaithersburg, MD: [s. n.], 1993: 173-187.
- [2] 李源, 郑毅, 何清, 等. 基于概念空间的文本语义索引 [J]. 计算机科学, 2002, 29(1): 20-22.
- [3] Voorhees E. Query expansion using lexical-semantic rela-

tions [C] // In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Dublin, Ireland: [s. n.], 1994: 61-69.

- [4] 冯兰萍. 本体在智能信息检索系统中的应用研究 [D]. 常州: 河海大学, 2005.
- [5] 王进. 基于本体的语义信息检索研究 [D]. 合肥: 中国科学技术大学, 2006.