基于本体视图特征项抽取方法研究 *

肖 升[1,2], 胡金柱[1], 姚双云[1], 舒江波[1]
(1. 华中师范大学 计算机科学系, 武汉 430079; 2. 湖南省第一师范学院 信息技术系, 长沙 410002)

摘 要: 为提供比单纯词汇信息更高效的概念特征信息和深层语义信息,并满足面向同一文本的多检索需求,在半自动化智能检索框架中引入本体视图,提出一种基于本体视图的特征项抽取方法。此方法首先针对文本特征建立本体视图;然后结合文本信息进行特征项抽取和类型映射,得到特征项集;最后基于特征项集完成检索过程。检索结果显示,基于本体视图特征项抽取方法能改善检索系统的性能,提高检索的准确率和效率。

关键词: 本体视图; 特征项抽取; 智能检索
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0042-03
doi:10.3969/j.issn.1001-3695.2010.01.011

Study on feature item extraction method based on ontology view

XIAO Sheng[1,2], HU Jin-zhu[1], YAO Shuang-yun[1], SHU Jiang-bo[1]
(1. *Dept. of Computer Science, Central China Normal University, Wuhan* 430079, *China*; 2. *Dept. of Information Technology, Hunan First Normal College, Changsha* 410002, *China*)

Abstract: This paper proposed a feature item extraction method based on ontology view by the introduction of ontology view in semi-automatic intelligent information retrieval framework. In order to provide concept feature and implied semantic which were more efficient than vocabulary information, and satisfied the demand for multi-queries faced the same Chinese text. Firstly, this method generated ontology view for text characteristic, and then obtained the set of feature item by extracted and mapped feature items with Chinese text information. At last, finished the process of multi-queries based on the set of feature item. The results show that the feature item extraction method based on ontology view can improve the performance of retrieval systems and improve the precision and efficiency of intelligent retrieval.

Key words: ontology view; feature item extraction; intelligent retrieval

0 引言

随着 Internet 海量信息与用户专一需求之间的矛盾日益凸显,寻找有效快捷的信息查询方法已成为 Internet 应用的当务之急。目前,主流的查询方法是以浅层统计模型(如向量空间模型)为核心的文本过滤方法[1]。此方法虽然便于实现且不依赖具体领域和语言,但由于缺乏对文档的语义分析,无法挖掘文本深层次主题信息,更无法保证以此为核心的系统性能。为了更好地实现对用户需求及 Internet 信息的语义理解,基于语义网的智能检索方法孕育而生。其中,基于知网的中文信息结构抽取方法最为成熟,此方法可抽取中文文本的结构信息及特征项,并根据用户需求对特征项进行计算,还可设定检索程序得到检索结果[2~4]。但是,此方法在同类文本特征项抽取方面存在缺陷,这使所构建的知识库在应对多个相关检索需求时显得分散而孤立[5]。为解决这一问题,本文引入一种基于本体视图的特征项抽取方法。此方法在已有领域本体的支持下构造本体视图,并以此视图为基础完成同类文本的特征项抽取。此方法能有效改进原智能检索框架中的特征项抽取环节,并能更好地理解非结构化数据源信息,提高智能检索的准确率和效率。

1 本体视图的概念

本体视图是从一个或几个本体中抽取出来的属性集,与数据库中的视图相对应,本体库中只存放本体视图的定义,当本体中某个概念或属性改变时,本体视图也要随之改变[6]。

本体视图的形式化表示为 $\mathrm{OnAttr}=\{a_1,a_2,\cdots,a_n\}$,本体有 n 个属性,OnAttr 为本体的属性集。$\mathrm{OnView}=\langle b_1,b_2,\cdots,b_m\rangle$ 是一个视图,$b_i\in\mathrm{OnAttr}$,那么,由 OnAttr 可构造的本体视图集为 $\mathrm{OnViewSet}=\{\langle b_1,b_2,\cdots,b_m\rangle\mid b_i\in\mathrm{OnAttr},m\in N\}$,进一步扩展为键类型对得到 $\mathrm{OnView}=\langle(b_1,t_1),(b_2,t_2),\cdots,(b_m,t_m)\rangle$。其中:$t_i\in T$,$T=\{$string, float, int, date, time, currency $\cdots\}$,T 为数据类型集。例如一个学术会议征文本体,其属性有截稿时间、论文修改时间、汇款时间、会议召开时间、会议地点、联系方式、征文范围、会议名称、主办单位。从这个本体可以创建视图 ov1:=〈会议名称,会议召开时间,会议召开地点,主办单位〉和 ov2:=〈会议名称,截稿时间,征文范围〉等。

收稿日期: 2009-04-09; **修回日期**: 2009-05-28 **基金项目**: 国家重点实验室开放研究基金资助项目(SKLSE04-018); 国家教育部重点研究基地重大研究项目(07JJD740063); 湖北省科技攻关资助项目(2007AA101C49); 湖南省教育"十一·五"规划重点项目(XJK06AZC010); 湖南省第一师范学院科研课题(XYSO9N04)

作者简介: 肖升(1980-),男,湖南武冈人,讲师,博士研究生,主要研究方向为中文信息处理(xiaosheng@mail.ccnu.edu.cn);胡金柱(1947-),男,湖北宜昌人,博导,主要研究方向为中文信息处理、软件工程;姚双云(1976-),男,湖南邵阳人,副教授,博士,主要研究方向为汉语语法、中文信息处理;舒江波(1982-),男,湖北宜昌人,博士研究生,主要研究方向为中文信息处理.

2　半自动化智能检索流程

半自动化智能检索中,特定主题是根据用户的需求,根据经验抽象出来的[7]。例如有如下检索需求:“2007 年的软件工程方向的征文信息”,可以根据经验得出这是一个关于“学术会议征文”的检索需求,则特定主题可以抽象为“学术会议征文”。这是一个人工处理过程,所以称之为半自动化智能检索,其过程如图 1 所示。具体过程如下:首先通过搜索引擎根据某一特定主题进行搜索,得到原始网页集 IWPS,IWPS 通过富文本解析器解析后得到原始纯文本 IPTS。同时,对于相关本体,构造本体视图;对于原始需求,通过需求分解,得到需求集;然后对 IPTS,结合本体视图进行特征项抽取和类型映射,得到可计算的特征项集;最后对特征项集和需求集应用检索算法,得到最终的符合用户搜索主题检索结果。

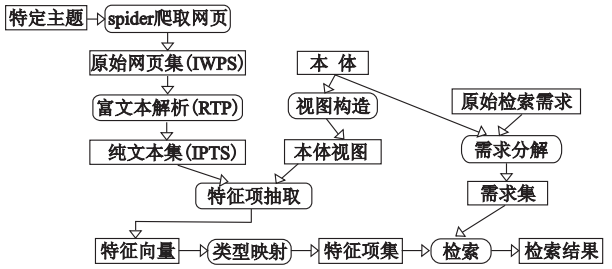

图1　引入本体视图的半自动化智能检索流程

3　基于本体视图的特征项抽取算法

3.1　算法步骤

a)文本表示成向量 $T=(T_1,T_2,\cdots,T_n)$,T_i 为中文信息结构分量。

b)计算本体视图的原像,即 $f^{-1}(\mathrm{OnView})$。其中:f^{-1} 是 f 的反函数,$f:T\rightarrow\mathrm{OnView}$。

记 $\mathrm{Dom}=f^{-1}(\mathrm{OnView})$。其中:$\mathrm{Dom}=\{X\mid X\subset T\wedge\forall x\in X, f(x)=\mathrm{OVA}\wedge\mathrm{OVA}\in\mathrm{OnView}\}$。

c)利用中文信息抽取器,从文本 T 中抽取 Dom 相关的信息,得到特征矩阵 FVM。

$$FVM=\begin{bmatrix}\langle X_{11},V_{11}\rangle & \langle X_{21},V_{21}\rangle & \cdots & \langle X_{n1},V_{n1}\rangle\\ \langle X_{12},V_{12}\rangle & \langle X_{22},V_{22}\rangle & \cdots & \langle X_{n2},V_{n2}\rangle\\ \text{ű} & \text{ű} & & \text{ű}\\ \langle X_{1m},V_{1m}\rangle & \langle X_{2m},V_{2m}\rangle & \cdots & \langle X_{nm},V_{nm}\rangle\end{bmatrix}$$

其中:n 为 OnView 的维数;m 是 Dom 中元素的最大特征数(称为本体视图属性的最大维)。每一列对应本体视图某一属性的相关信息,列中维数不足者补 0。

若 $f(X_1)=\mathrm{OVA}_1,|X_1|=m$ 为本体视图属性最大维;$f(X_2)=\mathrm{OVA}_2,|X_2|=n$,且 $n<m$,则 OVA_1 和 OVA_2 对应的矩阵列向量的转置分别为 $[\langle X_{11},V_{11}\rangle,\langle X_{21},V_{21}\rangle,\cdots,\langle X_{m1},V_{m1}\rangle]$ 和 $[\langle X_{12},V_{12}\rangle,\langle X_{22},V_{22}\rangle,\cdots,\langle X_{n2},V_{n2}\rangle,0,\cdots,0]$。

d)对特征矩阵的每一列进行概念消重、合并,得到特征向量 $E=(\langle\mathrm{OVA}_1,V_1\rangle,\langle\mathrm{OVA}_2,V_2\rangle,\cdots,\langle\mathrm{OVA}_n,V_n\rangle)$。其中 V_i 是 $(V_{i1},V_{i2},\cdots,V_{in})$ 消重、合并后的值。

e)类型映射 $f:E\rightarrow ME$,ME 是特征项集。$ME=\{\langle\mathrm{oid},\mathrm{aid},\mathrm{type},\mathrm{value}\rangle\}$。其中:oid 表示本体视图标志;aid 表示该视图的属性标志;type 表示该属性的数据类型;value 是可参与计算的值。四元组〈oid,aid,type,value〉称为一个特征项。例如特征向量 E 中的一个分量 $e=$〈截止日期,2007 年 5 月 30 日〉,则 $f(e)=\langle\mathrm{OV}_1,A_1,\mathrm{date},2007\text{-}5\text{-}30\rangle$。其中:$\mathrm{OV}_1$ 表示本体视图 1;A_1 表示属性 1;date 表示日期类型。

3.2　算法举例

为了使读者更清楚地了解这一过程,下面通过一个实例来简单说明。以“第四届中国软件工程大会征文通知”的文字片段为纯文本源,片段内容如下:“由浙江省信息产业厅主办,浙江省软件行业协会、希赛顾问团(CSAI)承办的第四届(2007)中国软件工程大会暨首届“天堂硅谷”中国软件产业高层人才论坛将于 2007 年 6 月 16 日至 17 日在美丽的西子湖畔召开……主办单位:浙江省信息产业厅、希赛顾问团(CSAI);承办单位:浙江省软件行业协会、希赛网;协办单位:《计算机教育》杂志社、浙江大学、杭州电子科技大学、杭州国家软件产业基地、湖南师范大学……,时间:2007 年 6 月 16 日至 17 日,地址:浙江省人民大会堂”。

a)文本表示成向量 $T=$(浙江省信息产业厅,主办,浙江省软件行业协会…,承办,…,2007 年 6 月 16 日至 17 日,西子湖畔,召开,…,主办单位,协办单位,承办单位)。

b)针对会议征文本体,取一个本体视图 OnView1 =(〈举办单位,string〉,〈会议召开时间,date〉,〈会议地点,string〉):f^{-1}(举办单位)=|主办,承办,协办,主办单位,承办单位,协办单位|;f^{-1}(会议召开时间)=|时间,于 $(date)召开|;$f^{-1}$(会议地点)=|地址,在 $(address)召开|;Dom $=f^{-1}$(OnView1)=|{|主办,承办,协办,主办单位,承办单位,协办单位|,|时间,于 $(date)召开|,|地址,在 $(address)召开|}。

c)利用中文信息抽取器,得到特征矩阵如下:

$$M=\begin{bmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ a_3 & 0 & 0\\ a_4 & 0 & 0\\ a_5 & 0 & 0\\ a_6 & 0 & 0\end{bmatrix}$$

其中:

$a_1=$〈主办,浙江省信息产业厅〉

$a_2=$〈承办,浙江省软件行业协会,希赛顾问团〉

$a_3=$〈协办,浙江大学,杭州电子科技大学……〉

$a_4=$〈主办单位,浙江省信息产业厅〉

$a_5=$〈承办单位,浙江省软件行业协会,希赛顾问团〉

$a_6=$〈协办单位,浙江大学,杭州电子科技大学……〉

$b_1=$〈时间,2007 年 6 月 16 日至 17 日〉

$b_2=$〈于 $(date)召开,2007 年 6 月 16 日至 17 日〉

$c_1=$〈地址,浙江省人民大会堂〉

$c_2=$〈在 $(address)召开,西子湖畔〉

d)对特征矩阵进行消重、合并。对第一列的“主办、承办、协办”进行合并,对第二列的“时间,于 $(date)召开”进行消重,对第三列的“地址,在 $(address)召开”进行消重合并,得到特征向量如下:

$E=$(〈举办单位,浙江省信息产业厅 & 浙江省软件行业协会 & 希赛顾问团 &《计算机教育》杂志社 & 浙江大学 & 杭州电子科技大学 &

杭州国家软件产业基地 & 湖南师范大学…〉,〈会议时间,2007 年 6 月 16 日 &2007 年 6 月 17 日〉,〈会议地点,西子湖畔 & 浙江省人民大会堂〉)

e)类型映射 f:$E\rightarrow ME$。

ME = {〈OnView1,举办单位,string,"浙江省信息产业厅 & 浙江省软件行业协会 & 希赛顾问团 & 浙江大学 & 杭州电子科技大学 & 杭州国家软件产业基地 & 湖南师范大学…"〉,〈OnView1,会议召开时间,date,"2007-6-16&2007-6-17"〉,〈OnView1,会议地点,String,"西子湖畔 & 浙江省人民大会堂"〉}

4　实验

根据以上方法,本文实现了一个基于本体视图的半自动化智能检索程序模块。对检索结果评价,最好是将程序得到的结果与实际的相关信息进行比较,观察不同的本体视图和检索算法对检索结果的影响。这需要一个完整的应用系统。由于是半自动化智能检索,采用人工判别的方法,对检索结果进行比较分析。本文选取"中国学术会议网"关于 2007 年的会议征文中的 211 条计算机科学类会议信息中的征文通知进行实验。其中,doc 文档有 27 篇,PDF 文档有 9 篇,部分会议没有征文通知,部分会议的征文通知是英文的,这些信息不在文本源选取范围内。按照半自动化智能检索流程,本文采用会议征文本体,其部分属性为 A:论文截稿时间,B:会议召开时间,C:主办单位,D:会议地点,E:会议名称。针对会议征文本体,建立了如下的本体视图:本体视图 OnView1 = 〈(A, date),(E, string)〉;本体视图 OnView2 = 〈(A, date),(C, string),(D, string),(E,string)〉。通过分解原始需求,得到如下需求集:检索需求 Q_1:2007 年 4 月份的会议;检索需求 Q_2:2007 年 4 月份在杭州召开的会议。说明:$Q\subseteq$OnAttr,Q 是需求集,OnAttr 是本体属性集。当 Q 的需求超出当前本体属性集,则可通过扩展本体,得到 OnAttr^+,使 $Q\subseteq\text{OnAttr}^+$。定义一个相关系数 r,r = 信息条数 / 检出信息条数。其中 r 越大,表明检索越准确,检出信息越接近目标。

OnView1 对应 Q_1,OnView2 对应 Q_2。需求分解算法和检索算法在其他文章中讨论,这里不详细给出其细节,只利用需求分解算法和检索算法进行知识处理。利用半自动化智能检索程序,得到如表 1 所示的结果。

表 1　基于本体视图的特征项抽取方法实验结果

	OnView1(A,E)	OnView2(A,C,D,E)
检出信息条数	9	3
相关	8	2
不相关	1	1
准确率	0.888 9	0.666 7
相关系数	3	9

通过对征文通知进行分析,发现在一个征文通知中有如下信息"征文截止日期:2007 年 3 月 15 日,修改稿截止日期:2007 年 4 月 5 日",由于本体视图的属性 A 是论文截稿时间,且映射为 f^{-1}(截止时间) = 论文截稿时间,在特征项抽取时得到的特征项为征文截止时间和修改稿截止时间的并,这导致了误差的出现。

5　结束语

本文提出了一种基于本体视图的特征项抽取算法,其基本思想是:针对本体建立本体视图,然后结合纯文本信息集进行特征项抽取以及类型映射,得到特征项集,利用特征项集对需求集进行检索,得到符合用户需求的结果。该方法能够从语义信息角度更好地表达文本内容,化简文本表示,对概念进行消重,有效地提高了智能检索的准确率。结合该方法之后的智能检索系统在处理非结构化数据上的实验结果也充分证实了该方法的有效性,为实现基于本体的自动化智能检索以及更好地利用纯文本中的语义信息提供了新的思路。

参考文献:

[1] 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,2003,34(15):109-111.

[2] 赵林,胡恬,黄萱菁,等. 基于知网的概念特征抽取方法[J]. 通信学报,2004,25(7):46-54.

[3] 董强,郝长伶,董振东. 基于知网的中文信息结构抽取[EB/OL]. (2005-11-10)[2006-04-12]. http://www.keenage.com.

[4] 刘群,李素建. 基于知网的词汇语义相似度计算[EB/OL]. (2005-10-18)[2006-05-17]. http://www.keenage.com.

[5] 秦进,陈笑蓉,汪维家,等. 文本分类中的特征抽取[J]. 计算机应用,2003,23(2):45-46.

[6] 唐杰,梁邦勇,李涓子,等. 语义 Web 中的本体自动映射[J]. 计算机学报,2006,29(11):1956-1976.

[7] 杨芳,杨振山. 基于语义网技术的主题词自动标引[J]. 计算机工程与设计,2005,26(10):2837-2839.

(上接第 41 页) 降低而不降低算法性能,并成功地应用于飞行控制系统参数寻优是进一步研究的问题。

参考文献:

[1] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proc of IEEE International Conference on Neural Netwoks. Perth: IEEE Press, 1995:1942-1948.

[2] 郝柏林. 从抛物线谈起:混沌动力学引论[M]. 上海:上海科技教育出版社,1993.

[3] ASANGA R, SAMAN K. Self-organizing hierarchical particle swarm optimizer with time-varing acceleration coefficients[J]. IEEE Trans on Evolutionary Computation, 2004,18(3):240-255.

[4] BERGH F van den. An analysis of particle swarm optimizers[D]. South Africa: Department of Computer Science, University of Pretoria, 2002.

[5] LIU Sheng-song, HOU Zhi-jian. Weighted gradient direction based chaos optimization algorithm for nonlinear programming problem [C]//Proc of the 4th World Congress on Intelligent Control and Automation. 2002:1779-1783.

[6] 吕振肃,侯志荣. 自适应变异的粒子群优化算法[J]. 电子学报,2004,32(3): 416-420.

[7] 孟红纪,郑鹏,梅国晖,等. 基于混沌序列的粒子群优化算法[J]. 控制与决策, 2006,21(3):263-266.

[8] 刘洪波,王秀坤,谭国真. 粒子群优化算法的收敛性分析及其混沌改进算法[J]. 控制与决策, 2006,21(6):636-640,645.

[9] KANG Qi, WANG Lei, WU Qi-di. Rearch on fuzzy adaptive optimization strategy of particle swarm algorithm[J]. International Journal of Information Technology, 2006,12(3):65-77.

[10] 段其昌,张红雷. 基于搜索空间可调的自适应粒子群优化算法与仿真[J]. 控制与决策,2008,23(10):1192-1195.

[11] 谷海红,齐名军,许少华. 一种基于混沌优化机制的双粒子群优化算法[J]. 计算机应用与软件, 2008,25(10):258-260.

[12] BILAL A, ERHAN A. Chaos embedded particle swarm optimization algorithms[J]. Chaos, Solitons & Fractals, 2009,40(3):1715-1734.