概念树在短文本语义相似度上的应用

赵小谦,郑 彦,储海庆
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:随着网络的发展,短文本语言计算的研究方兴未艾,且语义相似度在人工智能、认知学、语义学、心理学和生物学等领域里占有重要位置。在已有的传统的相似度研究算法上,为了能更快更准确地计算出相似度,文中通过构建概念树,设法把短文本集中到某个特定的领域。因概念树、概念词典既能表现概念之间的语义关系,又能表现概念层次结构,故而更能大大提高检索效率。在此基础上的相似度计算也使得检索结果更加准确,进而方便研究短文本之间的相似性与唯一性,大大增加了后期对挖掘的正确性。

关键词:短文本;WordNet;概念树;语义相似度
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2012)06-0159-04

Application of Concept Tree in Semantic Similarity of Short Texts

ZHAO Xiao-qian, ZHENG Yan, CHU Hai-qing
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: With the development of the network, short texts have attracted numerous researchers' attention, semantic similarity occupies an important positions in artificial intelligence, cognitive linguistics, semantics, psychology and biology. It is different from traditional essays on the research of semantic similarity, which tries to put the short text focus on some special area by building the concept tree. It's concept tree that shows the relationship and hierarchical structure between concepts, which more greatly improve the efficiency of searching, so as the concepts dictionary. On the basis of the similarity calculation makes the retrieval results more accurate, so it's more convenient to study the similarity and the uniqueness in short texts and the late mining.

Key words: short texts; WordNet; concept tree; semantic similarity

0 引 言

信息时代的到来,我国互联网事业和通讯事业发生了翻天覆地的变化,导致以电子形式存储和处理的数据爆炸性增长。这些数据中有很大一部分是长度很短的文本数据,且涉及的领域之广,深刻改变了亿万中国人的沟通方式和生活习惯。正因如此,对短文本的相似度计算,必须要分领域考虑。

现在使用的通用词典不能做到面面俱到,只在有的应用上显得非常成熟。如今的词汇量日益剧增,这就要求不能单单使用关键词匹配原则[1]进行匹配检索。与传统的词典不同,现如今的词典并不是仅仅要求把概念中的简单词汇按照某种序列编排起来。自然界中的事务之间都相互存在着联系,同样,概念也是如此。所以不仅要研究概念的含义,更多的时候,需要研究概念之间的相互关系,要把这些关系尽可能地在词典中体现出来。从而构建一个简单实用,且能方便地表示出概念与概念之间的联系关系,成了短文本相似度研究的关键。

收稿日期:2011-11-14;**修回日期**:2012-02-20
基金项目:国家重点基础研究发展规划(973)课题(2006AA01Z201)
作者简介:赵小谦(1986-),女,硕士研究生,研究方向为数据仓库与决策支持系统;郑 彦,教授,研究方向为数据仓库与决策支持系统。

1 概念间的相互关系

WordNet是最流行的英文语义词典,它也可以被看做是一个关于自然语言词条的一个本体[2,3]。它包含了约10万个词条,每个词条与一个或多个意思(一词多义)相对应。词条自顶而下被组织成分层的树状结构,靠近顶端的词条表示较广泛的概念,较低层次的词条表示较细致的概念。WordNet主要包含了名词、动词、形容词和副词这4大类词,词与词之间通过不同的关系相联系。其中最常用的关系是"是什么"关系和"整体—部分"关系,因为轮胎是汽车的一部分。通过这样的一些关系,词与词被联系了起来,不再是孤立的。HowNet[4]是一个在线的应用广泛的中文词语词典。HowNet不仅包含了词与词之间的关系,还将各个中文词与它们对应的英文词或解释联系起来。HowNet中包含的关系主要有上下位词关系、同义词关系、

部分—整体关系、相反关系、材料—产品关系、动态角色等等。最新的 HowNet 版本涵盖了约 11 万个概念。

在这里,以 WordNet 为例,由于所有的词都被自顶向下组织成了一个树状的结构,所以任意一个词都可以通过它们之间的路径长度和它们各自的深度信息反映出来。

上下位词关系如图 1 所示。

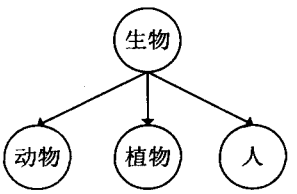

图 1 上下位词关系

2 概念树

2.1 概念树的结构

用语义网络描述概念之间的相互关系。根据数据结构知识可知,可以用概念树[5]方便的表示出语义网路。其中,领域概念的总集合可以用根节点来表示,最(较)小的概念用叶子节点表示。上层概念为子概念的概括,相反,子概念则为父概念的细分。

2.2 概念树的构建原则:

(1)手动给出概念树的层次框架,而框架的优化和完善则可以交给程序半自动来完成;

(2)父概念是子概念的总结,子概念是父概念的细分;

(3)不同的子概念代表父概念中的不同领域;

(4)子女可能不只有一个父母,即有向无环图可能成为概念树的层次结构。

2.3 构建概念树

由于概念关系很复杂,可以放在网络拓扑结构中研究,但是这无疑增加了难度和复杂度,为此,把上/下位关系作为树节点中'父/子关系',而其他的关系以关系指针的方式进行关联。因为研究树的特点往往要比网络容易的多。对于词典而言,最基本的操作就是检索,一部好的词典必须拥有相当高的查找效率,所以为此构造了索引表,来提高搜索效率,其原型为单链表结构,对应到概念节点的编码值[6]。

编码这个概念应该算是很常见的,商场物品有条形码,每个人有身份证号码等等。现在,从概念的词性、类别以及在词典中的结构三个角度对概念设计编码规则[7]。将编码分为两个部分:第一部分体现概念的含义,称之为"语义码";第二部分体现概念之间上/下位关系和在树中的位置,称之为"方位码"。接下来介绍语义码的构成。

(a) 语义码。定义词类编码映射表:对于不同的词性,用不同的符号进行标记,如表 1 所示:

表 1 词性编码映射表

词性	名词	动词	形容词	副词	介词	数量词	助词	连词	时间词	方位词
标记	n	v	a	d	p	q	u	c	t	f

其次,需对不同的类别进行进一步标记,使得概念之间区分得更加细致。以动词的 15 种类别为例进行标记,以 char s[5]存储其语义码,第 1 位确定它的基本词性,以'v'开头,后 4 位用以区别类别。例如身体动作动词(Verbs of Bodily Functions and Care),语义码为 vbody;通信动词(Verbs of Communication),语义码为 vcomm;变化动词(Verbs of Change),语义码为 vchng;竞争动词(Competition Verbs),语义码为 vcomp;消费动词(Consumption Verbs),语义码为 vcons 等等。

(b)方位码。顾名思义它能表现出概念在概念树中的具体位置,以及同其他概念之间的关系。由于根据概念间的相互关系进行词典的构建,所以,要在编码中能够体现出某些关系,比如上/下位关系。已经确立了建立概念树的方法,得知概念的上/下位关系通过树的'父子'关系来反映出来。

例如:有如下两个短文本,"小王和小张在交流最近各自的生活学习情况","小王和小张在说话",其中'交流'和'说话','说话'是'交流'的一种方式,它们之间形成上/下位关系,'交流'是'说话'的上位概念,反过来,'说话'即是'交流'的下位概念。假设'交流'在构建好的词典里的编码为 vcomm1321,其中'vcomm'是语义码,代表着通信类动词,'1321'称为方位码。以前缀码的方式来定义下位概念的编码,'说话'也同属于通信类动词,所以还是以 vcomm 开始,其方位码为'13211'。可以看出'1321'是其方位码的前缀部分,而'1321'是上位概念'交流'的方位码,这样体现了它们之间的上/下位关系。若有另一概念编码为'vcomm13212',可以看出其前缀码也是'1321',那么可以肯定的是,它与'vcomm13211'同属于上位节点'交流'的下位概念,进一步也体现了它们之间的兄弟关系。特别地,根节点的方位码定义为'1'。编码结构如图 2 所示:

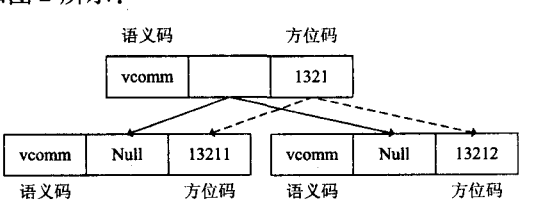

图 2 概念节点编码示意图

定义一张表格(如表 2),结合数据结构概念。当 Parent 为空时,则对应树的根节点。Parent 可以用来表示节点之间的上下位关系,例如 Parent = 1(A)则表示 B,C,D 的上位节点是 A,而 B,C,D 则为兄弟节点。其

中 Description 为对此概念进行描述。

表 2　上下位关系记录表

Code	Name	Parent	Description
1	A	Null	Null
11	B	1(A)	Nul
12	C	1(A)	Nul
13	D	1(A)	Nul
131	E	13(D)	Nul
132	F	13(D)	Nul
1321	G	132(F)	Nul

2.4　概念树构建算法

```
InPut: DataSource.
OutPut: Tree.
Tree(DataSouce)
{
Enter = SourceParse (DataSource);  //对源数据进行解析
TreeRoot = SearchRootNode(Enter);  //寻找根节点
BuildTree(TreeRoot); //构建概念树
//如果记录为空则跳出循环,否则在概念树中加入该列表
While(! Enter. Empty()){
FirstRecord=ReadTirstRecord();
//建立概念的上/下位关系
If (SearchParents(FirstRecord)) //是否存在该节点的父节点
{ AddedToTree(FirstRecord); }
else
{
AddedToTree(ParentTree);
//看是否有该节点的父节点,若没,则加入
}
//在已经构建完成的概念树中设置概念的其他关系指针
SetRalations(FirstRecord);
Records. Next();//移动到下一条记录
}
}
```

3　相似度分析

短文本相似度的计算方法主要分为以下几类:基于语义词典的方法[8]、基于大规模文本集进行统计的方法[9]、基于描述特征的方法[10]、借助互联网资源[11]的方法等。但是,文中主要涉及概念树在短文本语义相似度上的应用。所以,提出了基于概念树的短文本语义相似度的计算方法。

基本概念如下:

(1)构造一棵概念树,进行相似度计算。如图 3 所示 。

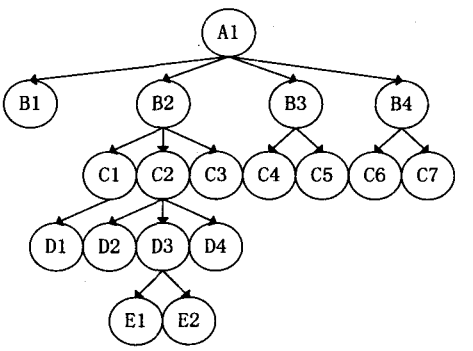

图 3　概念树

由数据结构基本知识可知:

父节点-----Parent(X):X 的直接上位节点,图 3 中的节点 D3 就是 E1,E2 的父节点。

祖先节点-----Ancestor(X):向上查找节点直至查找到根节点。图 3 中 E2 的祖先节点集为{A1,B2,C2,D3}。

共同祖先节点-----CoAncestors(X,Y):X,Y 共同的祖先节点集合。图 3 中 E1,D3 的共同祖先节点{C2,B2,A1}。

最近共同祖先节点-----NearAncestors(X,Y),即距离两个子节点距离最近的共同祖先节点。图 3 中 E1,D3 的最近共同祖先节点为 C2。

子树高度-----SubHeight(X):以根节点为起点,过 X 的最大路径长度。

概念深度-----Depth(X)= Depth(Parent(X))+1,其中,设 Root 为根节点,其概念深度为 1。

(2)语义距离。

语义距离是即语义树上概念与概念之间的路径,可以用它表示不同概念之间的相似程度[12]。长度语义距离和相似度之间成反比,语义距离越大,则相似度就越小。反之,亦然。构造一个函数,表示语义距离和相似度的关系:

$$e^{-\alpha Distance(X,Y)} \tag{1}$$

其中,$\alpha>0$ 为常数,表示语义距离和对于相似度的影响系数。

由上面所述的数据结构的概念得知,概念之间的相似度是由最近共同祖先节点表示[13]。其描述的越具体或越细致,则概念之间的相似度越大。从而得出这么一个函数:

$$\beta Density(NearConAncestor(X,y)) + \gamma \frac{Depth(NearCoAncestor(X,Y))}{Height(NearCoAncestor(X,Y))} \tag{2}$$

结合语义距离和共同祖先节点的知识，即公式(1)和(2)，进一步得出相似度函数(Similar Function)：

$$F(X,Y)= e^{-\alpha Distance(X,Y)} + \beta Desity(NearConAncestor(X,Y)) + \gamma \frac{Depth(NearConAncestor(X,Y))}{Height(NearCoAncestor(X,Y))} \quad (3)$$

其中 $\alpha+\beta+\gamma=1$，$F(X,Y)$ 的值域为(0,1]。

对于相似度计算公式(3)而言，假设，取 $\alpha=0.2$，$\beta=0.2$，$\gamma=0.2$，对图 3 的节点 E1 进行相似度计算，结果如表 3 所示：

表 3 各节点与 E1 的相似度计算

序号	查询节点	节点名称	相似度
1	E1	B1	0.3654
2	E1	C2	0.7467
3	E1	C4	0.5643
4	E1	D3	0.8675
5	E1	E2	0.9102

4 结束语

由此看出，可以根据构建出来的概念树，来分析概念与概念之间的上下位关系，再通过相似度的计算方法，用具体的数值表示出概念间的关系。从而可以针对某一个特定的领域来研究短文本，更便于找出短文本的相似性和唯一性。

文中虽然可以把短文本局限在某个特定的领域，但是，短文本的数量众多，且长度都非常短，样本特征非常稀疏，且词性灵活多变。所以，对于短文本相似度的研究，还存在众多难点，例如：怎样对短文本提取有效的特征词，怎样灵活处理短文本的词性问题，怎样更为周全的考虑到影响相似度的因素。这些都要进行下一步的研究。

参考文献：

[1] 邹文科. 基于本体技术的语义检索及其语义相似度研究[D]. 北京：北京邮电大学，2008.

[2] 黄 果，周竹荣. 基于领域本体的概念语义相似度计算研究[J]. 计算机工程与设计，2007，28(10)：2460-2463.

[3] 张凯勇. 基于 WordNet 的词语及短文本语义相似度算法研究[D]. 长春：吉林大学，2008.

[4] 贾文娟，何 丰. 基于 HowNet 的中文本体学习方法的研究[J]. 计算机技术与发展，2011，21(6)：120-125.

[5] 江 磊. 基于概念树的语义相似度计算的研究[J]. 广东通信技术，2010，30(8)：124-131.

[6] 樊兴华，孙茂松. 一种高性能的两类中文文本分类方法[J]. 计算机学报，2006，29(1)：124-131.

[7] 王洪伟，吴家春. 本体的形式化模型及在语义查询中的应用[M]. 北京：高等教育出版社，2003.

[8] 冉 婕，孙 瑜. 语义检索中的词语相似度计算[J]. 计算机技术与发展，2011，21(4)：40-48.

[9] 郑 诚，秦多荣. 本体映射中的概念相似度计算[J]. 计算机技术与发展，2008，18(11)：120-126.

[10] 谢信喜，王士同. 适用于数据的基于相互距离的相似性传播聚类[J]. 计算机应用，2008，28(6)：120-124.

[11] Montejo-Raez A，Dallman D. Experience in Automatic Keywording of Particle Physics Literature[J]. High Energy Physics Libraries Webzine，2001(5)：124-131.

[12] Zelikovitz S，Marquez F. Transductive Learning for Short-text Classification Problems Using Latent Semantic Indexing[J]. International Journal of Pattern Recognition and Artificial Intelligence，2005，19(2)：143-163.

[13] Sinha R，Mihaleca R. Unsupervised graph-based word sense disambiguation using measure of word semantic similarity [C]//Proceeding of the International Conference on Semantic Computing (ICSC07). Washington，DC，USA：IEEE Computer Society，2007.

(上接第 158 页)

参考文献：

[1] Guo Liang. Application of Improved Multivariate Linear Regression to Output Prediction of an Oilfield[J]. Journal of Xidian University (Social Science Edition)，2009，19(3)：71-75.

[2] 王瑞兵，范柱国. 滇中红层区滑坡灾害多元线性回归模型构建[D]. 昆明：昆明理工大学，2010.

[3] 董 冬. 缺失数据下线性模型回归系数岭估计的大样本性质[D]. 桂林：广西师范大学，2010.

[4] 林升梁. 多元线性回归模型在骨龄评估中的应用[J]. 吉林医学，2011，32(24)：5107-5108.

[5] Sun Baoqin. The Application and Strategy of Condition Based Maintenance of Power Equipment [J]. the Journal of Jilin Chemical Institute，2007，24(1)：21-25.

[6] 安徽省电力公司. 变电站内 35kV 及以下电压等级变压器类设备状态检修导则(试行)[S]. 2009.

[7] 安徽省电力公司. 变电站内 35kV 及以下电压等级变压器类设备状态评价导则(试行)[S]. 2009.

[8] Mathews J H，Fink K D. Numerical Methods Using MALTAB [M]. 4th ed. 北京：电子工业出版社，2009.

[9] 陈永胜，宋立新. 多元线性回归模型以及 SPAA 软件求解[J]. 通化师范学院学报，2001，28(12)：8-9.

[10] 田 兵. 多元线性回归分析及其实际应用[J]. 阴山学刊，2011，25(1)：16-19.

[11] 张凤莲. 多元线性回归中多重共线性问题的解决办法探讨[D]. 广州：华南理工大学，2010.

[12] 杨永生. 基于状态监测的机械设备可靠性评估模型[J]. 四川兵工学报，2010(7)：49-52.