文章编号:1001 - 9081(2008)11 - 2837 - 03

基于知网的概念匹配细粒度化研究

杨喜权[1],代 书[2]
(1. 东北师范大学 计算机学院,长春 130117; 2. 中国农业发展银行 科尔沁区支行,内蒙古 通辽 028000)
(yangxq375@nenu.edu.cn)

摘 要:基于知网的语义结构,构建了具有添加和删除特点的语义树,使概念的匹配粒度实现细化,并给出了概念语义树匹配算法。实验结果证明了算法的有效性,较好地解决"关键字障碍"和语义歧义性问题,提高查全率。

关键词:概念匹配;语义树;知网
中图分类号: TP311.13 **文献标志码**:A

Study on granularity concept matching based on How-Net semantic tree

YANG Xi-quan[1], DAI Shu[2]
(1. *School of Computer Science, Northeast Normal University, Changchun Jinlin* 130117, *China*;
2. *Horqin Borough Subbranch, Agricultural Development Bank of China, Tongliao Nei Mongol* 028000, *China*)

Abstract: In this paper, semantic trees that had accession and deletion were established based on How-Net to implement granularity concept matching. The concept semantic tree matching algorithm was presented. Efficiency of the algorithm was proved by the results of the experiment and the problem of "key word obstacle" and semantic ambiguity can be solved much better. The recall ratio was improved.

Key words: concept matching; semantic tree; How-Net

0 引言

在文本聚类中,短文档聚类存在高维稀疏性问题,导致了文档的查全率低下,本文基于这一问题在知网结构下进行概念语义分析,提出节点关键词映射知网关键词的匹配法,解决高维稀疏性问题。传统方法是基于文本关键字的向量空间模型(Vector Space Model, VSM),用 m 个关键字构成文档向量 $\boldsymbol{D}_i = \{d_{1i}, d_{2i}, \cdots, d_{mi}\}$ 表示文档集中的一个文档,这种方法表现在向量空间应用矢量内积计算文本向量空间的相似度,把词语看成独立的元素,词语之间没有联系,不能明确表达文本语义内容。其次,语义的向量空间模型只是对文本中存在的词语进行匹配,忽略词语中的一词多义以及一个文本语义的多种表示方法。

本文通过知网的内容来构建概念语义树,进行概念语义分析,消除一词多义、一义多词及一个文本语义内容可以有多种表达方式等问题的歧义性,并通过细粒度计算来解决稀疏性问题,提高查全率,从而将语义相近的文档实现基于内容的聚类本文。

1 语义树

语义分析可以分为含义确定,合成词语法分析,概念相似度计算三个主要部分。含义确定是指概念分析,在知网[1]中概念是由义原来描述的。概念用知识表示语言来描述,义原是描述概念的最小意义单位。知网并不是简单地将所有的概念归结到一个树状的概念层次体系中,而是试图用一系列的义原来对每一个概念进行描述。应用知网可以很好地把概念与其相关的词相结合,为构造概念语义树提供基础。

语义树[2]是按照某种规则在有向二叉树的每个节点上都标记有一个合式公式而构成。语义树采用树状的模型来构造语义空间,克服图状语义空间的不足。其特点表现在:1)构成语义树模型的元素是预先构造好的,但语义树本身是"即用即造"的,也就是根据特定的应用,动态灵活地构造语义树。2)动态构造的语义树可控性非常好,可以容易地添加或删除词。

1.1 构造概念语义树

语义树模型的构造有离线构造语义树和动态构造语义树。

1.1.1 离线构造语义树

设 $TSim(p,q)$ 为概念 p 与词 q 之间的相似度值,对任意给定的概念 p,采用树状的模型来表达概念 p 与所有其他概念的关系,如图 1 所示。

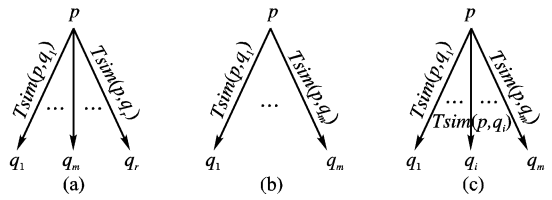

图 1 离线语义树

在图 1 中,以概念 p 为根节点,词 q 为叶节点,两者之间的路径权值是概念与词的相似度。从 q_1 到 q_r 的所有词都是按照词与概念 p 之间的相似度的大小来排序,以满足下式:$Tsim(p,q_1) \geqslant Tsim(p,q_2) \geqslant \cdots \geqslant Tsim(p,q_i) \geqslant \cdots \geqslant Tsim(p,q_m)$。

在图 1 三棵树中,确定相似度的方法是首先,保留了从左起到 m 个词,其余的词丢弃;其次从 q_1 到 q_m 的相似度区间为

收稿日期:2008 - 05 - 29;**修回日期**:2008 - 07 - 24。 **基金项目**:国家自然科学基金资助项目(60473042)。

作者简介:杨喜权(1963 -),男,吉林四平人,副教授,主要研究方向:知识发现与数据挖掘、语义 Web 与本体; 代书(1982 -),女,内蒙古通辽人,硕士研究生,主要研究方向:语义 Web。

(1 ~ 0.4),对保留的 m 个词进行再次划分,具体分为 q_1 到 q_i,q_i 到 q_m,定义 q_1 到 q_i 的相似度区间为(1 ~ 0.6),q_i 到 q_m 的相似度区间为(0.6 ~ 0.4)。这样可以更进一步地对语义进行比较,提高文本与文本集的细粒度聚类效果。

1.1.2 *动态构造语义树*

设 $\boldsymbol{P}=(p_1,p_2,\cdots,p_i,\cdots,p_n)$ 表示初始的义原组向量,其中 n 的值是随机的,p_i 为第 i 个义原组。通过 $\boldsymbol{P}$ 这个初始义原组向量构造语义树(Concept Similarity Tree Model, CSTM),$CSTM(\boldsymbol{P},v,j)$ 中的 v 代表语义树的层数,j 代表语义树每个元素(其中还包含 i-best 和 m-best),也就是每个相对根节点都连接着至少 m 个叶节点来构造语义树。如图 2 所示。

这样语义树由多个 i-best 树和 m-best 树在不同层次上构建起来,通过这棵语义树可以容易获得根节点概念与叶节点词之间的相似度值。具体定义如下:

1) 根节点 p 与它的叶节点 q 之间的路径是指节点 p 到节点 q 的路由。

2) 根节点 p 与它的叶节点 q 之间的路径的权值是所有在路径上的树枝的权值相乘,权值在构造 i-best 树和 m-best 树时给定。

3) 根节点 p 和它的叶节点 q 之间的最短路径是指他们的之间的最大权值的路径。

4) 根节点 p 和它的叶节点 q 之间的相似度是指他们之间的最短路径的权值。

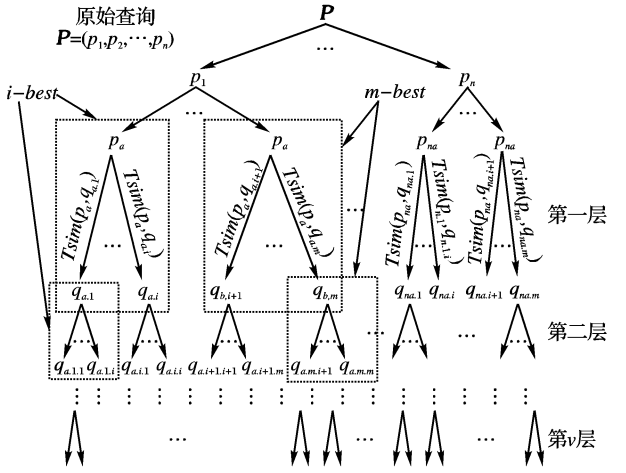

图 2 动态语义树 CSTM

在概念语义树中,路径越短,越靠近左侧说明它们越相似。P 为一个概念集,q 为词汇集,通常多个词汇可以映射到一个概念下,但是概念对词汇解释的程度是不同的,所以出现了概念与词之间的相似度问题。文献[3] 给出了式(1) 来计算相似度以便查询。设用户原始的主题概念向量是 $\boldsymbol{P}=(p_1,p_2,\cdots,p_n)$ 共包括 n 个概念,每个概念下连接的是关键主题词,关键主题词的向量为 $\boldsymbol{q}=(q_1,q_2,\cdots,q_i,\cdots,q_k)$,包含 k 个词,q_i 表示的是第 i 个词。当词 w 满足以下条件时,认为 w 是符合要求的扩展查询词。

$$\begin{cases} sim(\boldsymbol{q},w)=\sum_{i=1}^{k} sim(q_i,w)\geqslant cv \\ overlay(CSTM(\boldsymbol{p},v,j),w)\geqslant percent\times k \end{cases} \tag{1}$$

$sim(\boldsymbol{q},w)$ 是关键主题词向量 $\boldsymbol{q}$ 和词 w 之间的相似度,$sim(q_i,w)$ 是词 q_i 和词 w 之间的相似度,cv 为相似度的阈值。$Overlay(CSTM(\boldsymbol{p},v,j),w)$ 的值是在语义树 $CSTM(\boldsymbol{p},v,j)$ 中出现词 w 子树的个数,$percent$ 为覆盖度的阈值。

2 语义相关度

语义相关度[4] 反映了词语的词法、句法、语义甚至用语等特点。对词语相关度影响最大的是语义相关度。定义相关度为一个[0,1]之间的实数。

定义 1 语义相关度是在句法分析中一个短语结构的两个词能够组成修饰关系、主谓关系、同指关系的程度。

定义 2 在知网中,设 w_1 和 w_2 为任意两个词,w_1 有 n 个义原 $s_{11},s_{12},\cdots,s_{1n}$;$w_2$ 有 m 个义原 $s_{21},s_{22}\cdots,s_{2m}$;如果存在 $s_{1i}=s_{2j},1\leqslant i\leqslant n,1\leqslant j\leqslant m$ 则 w_1 与 w_2 的相关度为 1。

两个词的相似度高,那么相关度也高,但是相关度高并不表示相似度高,每一个类都是由解释义原相关联,义原树是以上下关系构成义原的相似度。义原与解释义原的关系形成了义原的关联度,所以说语义相关度是由义原的相似度与关联度所决定。

概念是根据多个义原组成的义原项来解释的,每个概念主要分为 4 个部分义原,在义原的体系结构中,每个义原与其他树中的义原也会存在一定关系,因此义原体系结构增加了横向联系,这样就存在义原的横向关联扩展。

根据语义相关度,对每个义原进行划分,在纵向与横向两个方面求出相关度可以更加准确地描述概念间的关系。给出相关度与相似度的公式如下。

$$R(\boldsymbol{p},\boldsymbol{q})=\max\left[\eta_1\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}H_j(S_1,S_2)+\eta_2\left(1-\frac{d(t_i,t_j)}{D}\right)\right],$$
$$\eta_1+\eta_2=1 \tag{2}$$

$$sim(\boldsymbol{p},\boldsymbol{q})=\sum_{i=1}^{k}sim(\boldsymbol{p},q_i) \tag{3}$$

在式(2) 中 t_i,t_j 为义原项 S_i,S_j 的第一基本义原,D 为横向关联影响的深度,$H_j(S_1,S_2)$ 表示的 4 个部分义原,$\beta_i(1\leqslant i\leqslant 4)$ 是可调节的参数,且 $\beta_1+\beta_2+\beta_3+\beta_4=1,\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$ 是 H_1 到 H_4 对总体相似度的影响,只有 β_1 有较大的权值。$sim(\boldsymbol{p},q_i)$ 是概念 $\boldsymbol{p}$ 和词 q_i 之间的相似度。根据文献[5] 中的定义证明如下。

证明

1) 对 $\boldsymbol{p},\boldsymbol{q}$ 两个集合,$\beta\in\boldsymbol{p},\alpha\in\boldsymbol{q}$,如果 $\alpha\in\beta,\alpha=\beta,\beta$ 是 α 的祖先,那么可以说明 $\boldsymbol{p}$ 与 $\boldsymbol{q}$ 相似。

2) 定义 $\boldsymbol{p}$ 的属性为 $WS_i=(I_1,I_2,\cdots,I_n,O_1,O_2,\cdots,O_m)$,$\boldsymbol{p}$ 的值域为 PD_i,定义域为 PR_i,其中 $PD_i\supseteq(I_1,I_2,\cdots,I_n)$,$PR_i\subseteq(O_1,O_2,\cdots,O_m)$;对任意元素 $p_i,q_j,\forall i,j$,如果 $(i,j)\in p_i$,那么 $(i,j)\in q_j$,因此 p_i 与 q_j 语义相似。

3) 任意的元素 p_i 与 $q_j,\forall i,j$,如果 $i\in PD_i,j\in PR_i$,所以 $i\in PD_j,j\in PR_j$,则称 p_i 与 q_j 语义相似。

匹配度公式(Matching Measurement):

$$MM=\frac{tR(\boldsymbol{p},\boldsymbol{q})+(1-t)ksim(\boldsymbol{p},\boldsymbol{q})}{d} \tag{4}$$

其中 $t\in(0,1)$ 为查询过程中所定义的阈值,k 为语义树的路径长度($k\geqslant 10$),$R(\boldsymbol{p},\boldsymbol{q})$ 为相关度,$sim(\boldsymbol{p},\boldsymbol{q})$ 为相似度,式(4) 把相关度与相似度相结合而得出匹配度。

概念向量 $\boldsymbol{p}$ 与词向量 $\boldsymbol{q}$ 通过相关度和相似度计算。在查

询文档时,把文档设为 W,W 的文档向量为$(w_1,w_2,\cdots,w_k)$,根据式(2)对文档向量 W 与 q 进行计算,对其所得的相似度与相关度逐层进行对比。当高一层的相似度与相关度低于阈值,计算终止。当高于阈值时逐层向下计算,直至到设定阈值为止,依次路径为语义查询路径。说明语义路径所标记的节点为根节点所属的概念。这样,通过调节 CSTM 的阈值,可以实现对概念语义不同粒度的划分。而仅考虑概念相似度计算的 VSM,一方面缺乏语义相关度的度量,另一方面概念相似粒度不可调。CSTM 算法如下:

设 P 为概念向量$(P_1,P_2,\cdots,P_n)$

输入:q 为词向量$(q_1,q_2,\cdots,q_n)$

输出:Simvalue;Tempvalue

```
if 1≤ R(p, q_m) ≤0.4;1 ≤ Sim(p, q_m) ≤0.4 then
  return q_m
else
  if R(p, q_i) ≥ R(p, q_m) ≥0.4; 0.4 ≤ R(p, q_i) ∪ R(p, q_m) ≤1;
  else if Sim(p, q_i) ≥ Sim(p, q_m) ≥ 0.4;0.4 ≤ Sim(p, q_i) ∪
    Sim(p, q_m) ≤1;
    if R(p, q_m) ≥0.4;
      for (i = 0, i < sim; i ++)
        { Svalue += sim[i] }
    if Svalue ≥0.4
      {return Simvalue = Svalue}
    else if R(p, q_m) ≤0.4 then
      return FALSE
      if R(p, q_i) ≥0.6;
        for (i = 0, i < sim; i ++)
          {Svalue += sim[i]}
      if Sim(p, q_i) ≥0.6;
        {return Simvalue = Svalue}
      if (R(p, q_i) ≥0.6, Sim(p, q_i) ≥0.6)
        {tempvalue = rvalue; tempIndex = i; Simvalue = Svalue;
         Simvalue = tempIndex;
         }
      if R(p, q_i) ≥0.4, Svalue ≥0.4
        for simθ = tR + (1 - t) k Svalue
          if simθ ≥0.4 then
            return TRUE
          else FALSE
```

3　实验

本文实验硬件环境:CPU 酷睿 2T7100;内存 1 GB;硬盘 110 GB/5400 RPS;操作系统为 Windows XP Professional SP2;软件环境 XML parser:JDOM1.0;数据库管理系统 Microsoft SQL Server 2000 SP4;开发工具 C++6.0。

输入 xml 文档:

```
<?xml version = "1.0" encoding = "gb2312"? >
<root >
  学校
  <node >
    大学
    <name >
      师范大学
      <college >
        计算机学院
        <specialty > 网络 </specialty >
      </college >
    </name >
  </node >
</root >
```

给定实验参数 $\beta_1 = 0.48$,$\beta_2 = 0.25$,$\beta_3 = 0.2$,$\beta_4 = 0.07$,$\partial = 1.6$,当阈值为 0.4 和 0.55 时,语义树的路径为 5,所求得匹配度的值及 VSM 相似度的值的比较,如表 1。

从表 1 中可以看出,基于 VSM 概念相似度计算,其值低于匹配度的计算。基于本文的二维计算(相似度和匹配度)中,一方面提高概念相似度的值,更为重要的通过改变阈值可调节概念相似的粒度。如师范大学概念在阈值 0.4 时,相似概念有 6 个,在阈值为 0.55 时,只有师范学院与之相似,匹配度为 0.660。可见,当阈值为 0.4 时,所查询的结果匹配粒度细致,降低稀疏性,提高查全率高。当阈值为 0.55 时,匹配精度高,查询结果与所需更为接近。

表 1　给定参数下的相似度

节点关键词	知网匹配关键词	VSM 相似度	匹配度(阈值:0.4)	匹配度(阈值:0.55)
学校	大学	0.513	0.649	/
	高中	0.331	0.431	/
大学	学校	0.563	0.697	/
	高校	0.540	0.688	/
师范大学	师范学院	0.656	0.838	0.660
	北京师范大学	0.551	0.715	/
	吉林师范大学	0.532	0.679	/
	东北师范大学	0.520	0.669	/
	华东师范大学	0.587	0.669	/
	白城师范大学	0.350	0.453	/
	长春师范大学	0.319	0.416	/
计算机学院	计算机系	0.560	0.720	0.570
	计算机科学	0.476	0.604	/
网络	网络教程	0.680	0.866	0.681
	网络理论	0.680	0.862	0.676
	网络技术	0.650	0.828	0.651
	网络安全	0.630	0.798	/
	网络通信	0.634	0.798	/
	通信技术	0.312	0.408	/

4　结语

本文在 TSTM 的基础上提出了概念语义树算法,算法基于知网的概念结构粒度可调地计算概念的相似度与相关度,得出节点关键词与知网匹配关键词的匹配度,提高查全率,从而实现文本与文本集间的聚类。这种方法可以很好地对词进行解释,减小词语匹配稀疏性,解决同义词和多义词问题。

参考文献:

[1] LIU QUN, LI SU-JIAN. Word similarity computing based on How-Net[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59 - 76.

[2] 李舟军, 王兵山. 语义树方法及其可靠性和完备性[J]. 国防大学学报, 1994, 16(3): 49 - 53.

[3] 赵军, 金千里, 徐波. 面向文本检索的语义计算[J]. 计算机学报, 2005, 28(12): 2068 - 2078.

[4] 许云, 樊春忠, 张锋. 基于知网的语义相关度计算[J]. 北京理工大学学报, 2005, 25(5): 411 - 414.

[5] CUI JUN-TAO, LIU JIA-MAO, WU YU-JIN, *et al*. An ontology modeling method in semantic composition of Web services[C]// Proceedings of the 2004 IEEE International Conference on E-Commerce Technology for Dynamic E-Business. Washington, DC: IEEE Computer Society, 2004: 270 - 273.