

# 用于形式背景提取的中文文本表示

侯亚南, 黄映辉

(大连海事大学 信息科学技术学院, 辽宁 大连 116026)

**摘要:**形式背景需要从实际的数据源中提取。当数据源为无结构的中文文本时,必须选择如何对其进行表示。目前主流的中文文本表示方法主要采用以词语为特征项的向量空间模型(VSM),其主要缺陷是忽略了自然语言中词语之间的语义联系,无法表达文本的语义信息。讨论了一种改进方法,其特征是:选择知网(HowNet)作为知识库,采用相似词集集合代替单一特征词,建立中文文本的概念向量空间。对于用概念向量空间表示的中文文本,可以方便地根据用户的具体要求提取所需的形式背景。以214篇交通类中文文本为实例阐释了该改进方法的实际应用。

**关键词:**形式背景;文本表示;相似词集集合;向量空间模型

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2010)09-0036-04

## Chinese Document Representation for Extracting Formal Context

HOU Ya-nan, HUANG Ying-hui

(Information Science and Technology College, Dalian Maritime University, Dalian 116026, China)

**Abstract:** A formal context must be extracted from data sources. But to extract a formal context from unstructured Chinese document needs to decide how to represent it first. The dominant model of document representation, which is called the Vector Space Model (VSM), uses a single word as the characteristic item. It is obvious that VSM neglects the lexical semantic relation between words, thereby it can not express the semantic information of documents. Discusses an improved method which is to take HowNet as knowledge base, to establish the concept vector space of Chinese document by using the set of similar word set to replace the single characteristic item in VSM. On the base of Chinese document with concept vector space, it is convenient to extract the formal context to meet user demand. Illustrate the application of this improved method with 214 Chinese texts about transportation as examples.

**Key words:** formal context; document representation; set of similar word set; vector space model

## 0 引言

形式概念分析(formal concept analysis, FCA)是R. Wille提出的知识处理新方法<sup>[1]</sup>,其应用的第一步就是建立形式背景(formal context)。形式背景需要从实际的数据源中提取。文本(document)是一种最常用的数据源形式,可分为有结构(unstructured)和无结构(structured)两种<sup>[2]</sup>。

无结构文本(unstructured document)的内容大多是用自然语言表示的,计算机无法直接理解其语义从而不能进行有效地处理。因此,对于无结构文本需要进行预处理,抽取能表达文本主题内容的特征,并将这些特征用结构化的形式表示出来,即文本表示(document representation)<sup>[3]</sup>,以便机器能够对其做进一步处理。

目前主流的文本表示方法是向量空间模型(vector space model, VSM)。

以词语为特征项的向量空间模型的主要缺陷是忽略了自然语言中词语之间的语义联系,无法表达文本的语义信息。文中尝试对其进行改进,提出中文文本的相似词集集合(set of similar word set)表示方法。

考虑到中文知识库在中文词义处理上已经积累了一定的成果<sup>[4]</sup>,文中选择知网(HowNet)作为知识库,采用相似词集集合代替单一的特征词,建立概念向量空间以表示中文文本。通过将知网描述的概念引入到向量空间模型中,从结构和语义两个方面弥补向量空间模型在中文文本表示上的不足。

## 1 相关知识

### 1.1 形式背景

形式背景被定义为一个三元组  $K = (G, M, I)$ , 其中,  $G$  为对象集合,  $M$  为属性集合,  $I$  为  $G$  和  $M$  之间的二元关系, 即  $I \subseteq G \times M$ 。该三元组可以表示为二维

收稿日期:2010-01-19;修回日期:2010-04-04

基金项目:国家自然科学基金资助项目(60972090)

作者简介:侯亚南(1985-),女,山东济宁人,硕士研究生,研究方向为智能信息处理;黄映辉,教授,主要从事智能信息处理的研究。

表。在表 1 所示的形式背景中,对象集合  $G = \{f_1, f_2, f_3, f_4\}$ ,属性集合  $M = \{a, b, c, d\}$ ,二元关系  $I$  为确定性关系。

表 1 形式背景的示例

	a	b	c	d
$f_1$	1	1	0	0
$f_2$	1	0	1	1
$f_3$	0	1	1	1
$f_4$	0	0	1	1

1.2 知 网

知网(HowNet)是一个以词语形式表示的概念为描述对象、以揭示概念之间以及概念所具有的属性之间的关系为基本内容的通用知识库<sup>[5]</sup>,由多个数据文件构成。知网的基本思想是将各个词语的词义以义项(sense)表示,再通过标准化的义原(primitive)描述各个义项。知网对每一个词语,不仅标注了其语义和词性,而且标注了它所属的概念、概念与其内部属性之间的联系以及概念之间的联系。概念是事物本质特征的概括和抽象,且不受词语的语种、多义性和歧义的影响<sup>[6,7]</sup>。

知网采用知识描述语言(knowledge database markup language, KDML)<sup>[5,8]</sup>来表示词语所代表的概念。例如,对“高手”的 KDML 描述为“高手:DEF = {human|人, able|能力, desired|良好}”,通过三个义原所表达的清晰语义为“高手是一个能力优良的人”。

在知识描述语言中,义原是最基本的不可再分割的最小意义单位,一个词语可以表示多个概念(即一词多义)。此外,知网还定义了上位、下位、同义、反义等 16 种关系,这些关系描述了概念之间的相互联系,从而使知网成为一个网状的知识系统<sup>[9]</sup>。

1.3 向量空间模型

向量空间模型(VSM)由 Salton 等提出,其基本思想是:以词语作为特征项,将文本表示为由特征项构成的向量空间中的一个点,通过计算向量之间的距离判断文本之间的相似程度。文本空间中的每个文本都被表示为一个向量,向量的每一维对应一个特征<sup>[10]</sup>。

向量空间模型是一种基于统计的文本表示模型,根据词频(term frequency)和逆文本频率(inverse document frequency)来决定特征词的权重。由于不考虑词语之间的上下文关系,从而忽略了上下文关系所蕴涵的语义信息,这是该模型无法解决同义词与多义词问题的根本原殷。同时,在向量空间模型中词语处于散列、无结构的状态,无法表现出词语之间的关联关系,因而大大降低了其在实际应用中的表现能力。

1.4 相似词集集合

相似词集(similar word set)是词语之间的相似度

超过某一阈值的词语的集合。一个相似词集可用于表示一个概念。例如,“银行”、“在线银行”和“网上银行”可组成一个相似词集,其表示的概念是“借贷场所”;“好人”和“坏人”组成的相似词集可以表示概念“人”。

相似词集集合(set of similar word set)是用相似词集表示的概念的集合,是概念集合的一种表示方式。若概念集合是{借贷场所,人},则对应的相似词集集合就是{{银行,网上银行,在线银行},{好人,坏人}}。

相似词集内部的特征词之间是高内聚的,而相似词集之间则是低耦合的;同时,相似词集集合保证了同一个特征词若在文本中表示的词义不同、不会被认为是相同词而筛查剔除。

一个文本可以被表示为一个相似词集集合,该集合的元素即是文本的特征词经过词义消歧后所得到的相似词集。

2 中文文本的相似词集集合表示

采用相似词集集合表示中文文本的方法如图 1 所示。无结构的中文文本经过文本预处理、特征词义项预处理、特征词权重计算、词义消歧、词向量空间重构而成为用相似词集集合表示的结构化的中文文本。

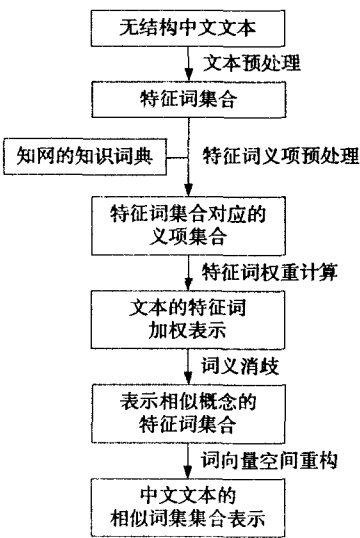


图 1 中文文本的相似词集集合表示

2.1 文本预处理

文本预处理是文本内容能被计算机识别的基础,其主要步骤如下:

- (1)去除文本中的非字符信息,例如图像等。
- (2)把文本划分表示成句子集合。将文本表示为:  $D = \{K_1, K_2, \dots, K_i, \dots, K_p\}$ ,其中  $K_i$  表示文本的第  $i$  个句子。
- (3)对句子进行分词处理。将句子表示为:  $K_i = \{W_{i1}, W_{i2}, \dots, W_{ij}, \dots, W_{ik}\}$ 。其中  $W_{ij}$  为句子  $K_i$  的第  $j$

个特征词。分词工具采用中国科学院的开源分词工具 ICTCLAS。

(4) 去除停用词,例如“的、了”等。

经过文本预处理后,文本表示为句子集合,句子表示为特征词集合。

## 2.2 特征词义项预处理

在知网上查询文本的特征词所对应的概念,即进行特征词义项预处理,其目的在于确定文本中任意特征词具有的词义(可能多个),并记录义项号。特征词义项预处理的过程是:

- (1) 依次遍历文本  $D$  的每个句子  $K_i$ ;
- (2) 依次遍历句子  $K_i$  的每个特征词  $W_{ij}$ ;
- (3) 查询知网,并记录  $W_{ij}$  对应的义项。

通过特征词义项预处理,得到文本的每个句子的每个特征词所对应的义项集合。例如,特征词“医生”所对应的义项集合是{098818}。

## 2.3 特征词权重计算

特征词权重的计算选择常用的 TF-IDF 方法<sup>[11]</sup>。特征词  $W_i$  的权重为:

$$W_{ID} = TF_{ID} \cdot IDF_i \quad (1)$$

其中,  $TF_{ID}$  为特征词  $W_i$  在文本  $D$  中的词频,  $IDF_i$  为特征词  $W_i$  在文本  $D$  中的逆文本频率。

## 2.4 词义消歧

一词多义迄今仍是中文处理的难题。相关研究表明词义消歧对机器翻译、信息检索、文本分析、自动文本分类等许多方面都有十分重要的作用<sup>[12]</sup>。文中以知网为基础,通过计算词语语义相似度来实现词义消歧,其基本思想是:计算待消歧词语的各词义与该词语所在上下文中其他词语的语义相似度,根据语义相似度值所反映的词义之间的关联关系实现多义词消歧。

文中采用利用知网计算词语语义相似度的方法<sup>[13]</sup>。义项  $S_1$  和义项  $S_2$  的语义相似度为:

$$\text{Sim}(S_1, S_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i \text{Sim}_j(p_1, p_2) \quad (2)$$

其中,  $\beta_i (1 \leq i \leq 4)$  是可调节的参数,且  $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$ ;  $\text{Sim}_j(p_1, p_2)$  是第  $j (1 \leq j \leq 4)$  层义项描述的义原  $p_1$  和  $p_2$  之间的语义相似度:

$$\text{Sim}_j(p_1, p_2) = \frac{\alpha}{\alpha + d} \quad (3)$$

其中,  $\alpha$  是一个可调节的参数;  $d$  是  $p_1$  和  $p_2$  在义原层中的路径长度;  $j$  是义项描述的层级,在知网中义项描述共分为四层,依次为:第一基本义原描述、其他基本义原描述、关系义原描述、关系符号描述。

中文词义消歧算法以句子为处理单位,即以句子作为词语所在的上下文。设句子  $K$  中有  $n$  个特征词

$W_i (1 \leq i \leq n)$ , 特征词  $W_i$  有  $N_i$  个义项  $C_{ij} (1 \leq j \leq N_i)$ , 则句子  $K$  中的特征词  $W_i$  的第  $j$  个义项  $C_{ij}$  与特征词  $W_k (k \neq i)$  的相似度为:

$$S(C_{ij}, W_k) = \max(\text{Sim}(C_{ij}, C_{kp})) \quad (4)$$

特征词  $W_i$  的义项  $C_{ij}$  的权重,即义项  $C_{ij}$  为所求义项的可能值为:

$$\text{Value}(C_{ij}) = \frac{\sum_{k=1, k \neq i}^n S(C_{ij}, W_k)}{n-1} \quad (5)$$

以上计算,综合考虑了特征词  $W_i$  所在的上下文和有利于词义消歧的多种因素。据此,文中提出的特征词  $W_i$  的词义消歧算法为:

For  $j = 1$  To  $j = N_i // N_i$  为义项个数

{

For  $k = 1$  To  $k = n // n$  为特征词个数

{

If  $k \neq i$  计算  $S(C_{ij}, W_k)$ ;

}

计算  $\text{Value}(C_{ij})$ ;

}

取最大  $\text{Value}(C_{ij})$  对应的义项;

由该算法求得的最大  $\text{Value}(C_{ij})$  对应的义项即为特征词  $W_i$  在当前语境中的词义。

## 2.5 词向量空间重构

文本经词义消歧后,其每个特征词都对应唯一的概念。词义消歧后的文本  $D$  可被表示为  $D = \{(W_1, C_1), (W_2, C_2), \dots, (W_i, C_i), \dots, (W_q, C_q)\}$ 。其中,  $W_i$  为特征词,  $C_i$  为  $W_i$  对应在知网中的义项号,即知网中描述的概念。

词向量空间重构的主要步骤为:

(1) 消除相同词。若式(5)中存在义项号相同的元素,则消除至仅保留一个。

(2) 构建相似词集。利用式(2)计算两个词语的语义相似度,设定阈值以确定相似词。

(3) 构建文本  $D$  的相似词集集合,即:

$$D = \{(W_{11}, C_{11}), (W_{12}, C_{12}), \dots, (W_{1h}, C_{1h})\}, \{(W_{21}, C_{21}), (W_{22}, C_{22}), \dots, (W_{2g}, C_{2g})\}, \dots, \{(W_{n1}, C_{n1}), (W_{n2}, C_{n2}), \dots, (W_{nf}, C_{nf})\} \quad (6)$$

文中相似词集的权重定义为该集合中所有相似词的权重之和。相似词集的权重越大,表示该集合与文本  $D$  的关系越密切。

综上所述,中文文本的相似词集集合表示,通过对无结构中文文本的预处理得到特征词集合列表,利用词语在知网中的描述将中文文本中具有一词多义的特征词进行词义消歧,同时对特征向量进行重构,使其内

部的特征词组织化、有序化。通过计算权重所得结果,制定选取条件得到符合用户需求的中文文本表示,在此基础上可以方便地提取形式背景。

3 提取形式背景

从文本中提取形式背景是根据用户需要制定阈值提取有用信息的过程。

对于有结构的文本,可直接通过程序自动提取,根据用户自定义的阈值转化为用户所需的形式背景。

对于无结构文本,必须首先将其转换成有结构的文本。文中所述方法是用相似词集集合表示中文文本。在此基础上就可以根据相似词集的权重以及用户自定义的阈值进行处理,从而提取获得形式背景。

对于相似词集集合表示的中文文本,所有文本作为形式背景对象集,相似词集集合作为属性集,权重值超过阈值取 1,否则取 0,即得用户所需的形式背景。

4 实例验证

实例选择搜狗实验室([http://www.sogou.com / labs](http://www.sogou.com/labs))提供的 10 大类 2816 篇中文常用文本中的 214 篇交通类中文文本。

4.1 中文文本的相似词集集合表示

将第 2 节所述方法应用于实验数据,过程如下:

(1)文本预处理。通过对 214 篇文本处理形成特征词列表,其中根据词频数控制特征词的提取,文中以去掉词频小于 10 或大于 30 的词为例,得到 314 个特征词。

(2)特征词义项预处理。通过在知网中查询义项号,得到与特征词列表对应的义项号列表。

(3)特征词权重计算。结果如表 2 所示。

表 2 实例的特征词权重

	车厢	驾驶员	司机	时速	速度
文本 1	0	0.04	0.12	0.26	0.45
文本 2	0.65	0.02	0.06	0.21	0.16
文本 3	0	0.22	0.40	0	0
文本 4	0.51	0.43	0.14	0.32	0.15

(4)词义消歧。由于实验结果数据较多,故仅以“速度与时速”和“驾驶员与司机”为例,如表 3 所示。

表 3 实例的词义消歧举例

	语义描述	相似度
速度	属性,速度,变空间位置	1.00
时速	属性,速度,变空间位置	
驾驶员	人,职位,驾驭,车	1.00
司机	人,职位,驾驭,车/飞行器	

(5)词向量空间重构。由表 2 和表 3 可得相似词集集合为{{车厢},{驾驶员,司机},{时速,速度}},再

计算各文本的相似词集的权重。这样就得到中文文本的相似词集集合表示,结果如表 4 所示。例如,文本 2 = {{车厢=0.65},{驾驶员,司机=0.08},{时速,速度=0.37}}。

表 4 文本的相似词集集合表示

车厢	驾驶员/司机	速度/时速	
文本 1	0	0.16	0.71
文本 2	0.65	0.08	0.37
文本 3	0	0.62	0
文本 4	0.51	0.57	0.47

4.2 提取形式背景

在表 4 所示的中文文本的相似词集集合表示的基础上提取形式背景。若选择权重大于 0.45(即阈值为 0.45)的特征词,则结果如表 5 所示。

表 5 实例所提取的形式背景

	车厢	驾驶员/司机	速度/时速
文本 1	0	0	1
文本 2	1	0	0
文本 3	0	1	0
文本 4	1	1	1

5 结束语

文中在空间向量模型的基础上引入知网描述的概念,其基本思想是利用知网获取词语的语义信息,将中文文本用相似词集集合表示。该方法能够从语义信息角度更好地表达文本内容,体现了词向量空间中词语之间的相关性。在词义消歧算法上,文中综合考虑了词语所在上下文、知网的结构以及语义相似度计算方法等方面对词义消歧的影响。通过实例验证表明,该中文文本表示方法是可行有效的。

参考文献:

[1] Cross V, Yi Wenting. Formal concept analysis for ontologies and their annotation files[J]. Fuzzy Systems, 2008(3):2014 - 2021.

[2] 韩道军, 张 磊, 沈夏炯, 等. 形式背景提取初探[J]. 河南大学学报, 2007, 37(5):523 - 526.

[3] 郭少友. 自动分类中的文档表示及其改善方法研究[J]. 信息技术, 2008(8):23 - 25.

[4] Leacock C, Chodorow M, Miller G. Using corpus statistics and WordNet relations for sense identification[J]. Computational Linguistics, 1998, 24(1):147 - 166.

[5] 董振东, 董 强. 知网简介[EB/OL]. 1999 - 01 - 01. <http://www.keenage.com/html/c-index.html>.

[6] 张 剑. 基于概念的文本表示模型的研究[D]. 北京: 清华大学, 2006.

[7] 郝长玲, 董 强. 知网知识库描述语言[C]//全国第七届计

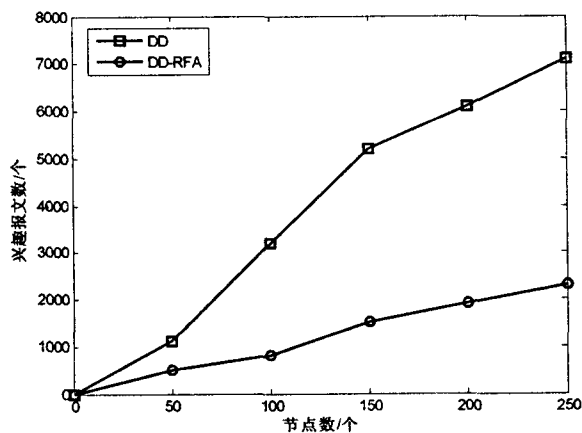


图4 不同网络环境下的兴趣报文数

如图5所示:改进的DD协议,有效地降低了节点的能量开销,这是由于利用射频角度的方法,使得部分不在正确传输方向上的节点不参与兴趣的转发。这样就大大节省了节点的能量,由于降低了节点能量开销,使得节点寿命得到了很好的延长,有利于延长整个网络的生命周期。

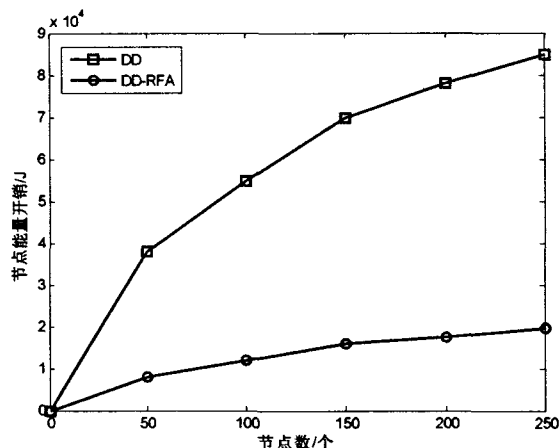


图5 不同网络环境下的能量开销

## 4 结束语

阐述了DD协议的基本原理,并针对DD协议泛洪兴趣所引起的巨大的能量消耗,提出了基于射频发射

角度的改进协议(DD-RFA)。DD-RFA很大程度上缩小了兴趣泛洪的范围,减少了参与泛洪节点数目,从而节省了节点能量,延长了整个网络的生命周期。

## 参考文献:

- [1] Akyildiz I F, Su W, Sankarasubramaniam Y, et al. A survey on sensor networks[J]. IEEE Communications Magazine, 2002, 40(8):102-114.
- [2] 李建中, 高宏. 无线传感器网络的研究进展[J]. 计算机研究与发展, 2008, 45(1):1-15.
- [3] Al-Karaki J N, Kamal A E. Routing techniques in wireless sensor networks: a survey[J]. IEEE Wireless Commun, 2004, 11(6):6-28.
- [4] 任丰原, 黄海宁, 林闯. 无线传感器网络[J]. 软件学报, 2003, 14(7):1248-1291.
- [5] 唐勇, 周明天, 张欣. 无线传感器网络路由协议研究进展[J]. 软件学报, 2006, 17(3):410-421.
- [6] Intanagonwiwat C, Govindan R, Estrin D. Directed diffusion: a scalable and robust communication paradigm for sensor networks[C]//In: Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking. [s.l.]:[s.n.], 2000: 56-67.
- [7] Intanagonwiwat C, Govindan R, Estrin D, et al. Directed diffusion for wireless sensor networking[J]. IEEE/ACM Transactions on Networking, 2003, 11(1):2-16.
- [8] Macro D, Maniezzo V, Colomi A. The ant system: optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics - Part B, 1996, 26(1): 29-41.
- [9] 苏均宇, 曾子维, 石嘉鹏. 基于定向扩散路由协议的改进[J]. 传感技术学报, 2007, 20(3):673-676.
- [10] The Network Simulator - ns-2[EB/OL]. 2008-03-31. <http://www.isi.edu/nsnam/ns/>.
- [11] 徐蕾鸣, 庞博, 赵耀. NS与网络模拟[M]. 北京:人民邮电出版社, 2003.
- [12] 柯志亨, 程荣祥, 邓德隽. NS2 仿真实验——多媒体和无线网络通信[M]. 北京:电子工业出版社, 2009.

(上接第39页)

- 算语言学联合学术会议论文集. 北京:清华大学出版社, 2003:371-377.
- [8] 张明宝, 马静. 一种基于知网的中文词义消歧算法[J]. 计算机技术与发展, 2009, 19(2):22-25.
  - [9] Salton G, Wong A, Yang S. A vector space model for automatic indexing[J]. Communication of ACM, 1975, 18(11): 613-620.
  - [10] Ide N, Veronis J. Introduction to the special issue on word sense disambiguation: the state of the art[J]. Computational Linguistics, 1998, 24(1):1-40.

- [11] 晋幼丽, 周成全, 王学松. SVM和K-means结合的文本分类方法研究[J]. 计算机技术与发展, 2009, 19(11):35-37.
- [12] Schutze H, Pedersen J. Information retrieval based on word senses[EB/OL]. 2007-12-01. <http://sholar.google.com.cn>.
- [13] 刘群, 李素建. 基于知网的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会. 台北:[出版者不详], 2002:1-18.