文章编号:1009-3087(2014)05-0121-06

基于可拓知识库的短文本敏感信息查询扩展方法

张海涛[1],王斌君[1],张洪涛[2]
(1. 中国人民公安大学,北京 100038;2. 哈尔滨市公安局,黑龙江 哈尔滨 150001)

摘 要:为了解决一般检测算法在短文本查询上效率较低的问题,基于可拓学的方法构建特定领域可拓知识库并定义一种新的菱形推理模式,提出了基于可拓知识库的概念查询扩展算法,通过概念间的可拓关联性及可拓推理来处理短文本中敏感信息检测问题,并通过实例验证了算法的可行性。实验证明该算法克服了短文本自身长度较短、描述概念能力弱的问题,可减少相关信息的遗漏,该算法提高了文档敏感信息检测的准确率与召回率。

关键词:短文本;可拓知识;敏感信息;查询扩展;知识库
中图分类号:TP393 **文献标志码**:A

A Query Expansion Method of Sensitive Information for Short Text Based on Extension Knowledge Base

ZHANG Haitao[1], *WANG Binjun*[1], *ZHANG Hongtao*[2]
(1. People's Public Security Univ. of China, Beijing 100038, China; 2. Harbin Public Security Bureau, Harbin 150001, China)

Abstract: In order to solve the problem of the low efficiency caused by the traditional query expansion retrieval methods, a novel approach based on extension knowledge was issued with a novel rhomb reasoning method by using extension theory. By introducing the extension relationship between concepts with extension reasoning, the proposed method can relationally solve the detection problem of sensitive information. For illustration, an example was utilized to show the feasibility of the method in solving detection problem with concept query expansion method. Empirical results showed that the proposed method has a good performance on detection of short text which has a low degree of description, and could reduce the omission of relative information. The proposed method can improve the detection precision and recall of the document sensitive information.

Key words: short text; extension knowledge; sensitive information; query expansion; knowledge base

互联网的深入应用,网上出现了多样化的、大量的短文本数据,比如网上论坛、博客、留言、评论等,其字数较少,却数量巨大,单条文本所含信息量小,但信息的真实性与及时性更强。从语言学角度看,此类信息具有语法不完整、描述不规则、文本冗余、歧义等不符合语言规范的特点。针对这些问题,如何有效地发掘和利用这些大量的短文本信息,筛选与提炼出针对性的敏感信息,已经成为舆情分析领域一个重要的课题。

现有的敏感信息查询扩展方法中,基于词语全局聚类的查询扩展技术[1]是较成熟的一种,原理是将全部文档按某种聚类方法生成不同的簇,然后分别为其构造相应的局部叙词表。该方法确实能提高检索性能,但在处理查询词的歧义性问题时性能欠佳。

查询词权重调整方法[2]是从经过人工选出的初始文档中按照需要选择欲加大权重语词,将其添加到原查询序列中组成新的查询词。反馈信息能够用于对原查询语词的重新加权,提高了检索的召回率,但增加了较多人工因素,因为相关性的判断是取决于相关经验。

基于局部上下文分析的(LCA)查询扩展技术[3-4]

收稿日期:2014-03-07
基金项目:中央高校基本科研业务费专项资金资助项目(2013LGX02)
作者简介:张海涛(1982—),男,博士生. 研究方向:信息安全;计算机犯罪侦查. E-mail:okhaitao@126.com
网络出版时间:2014-7-10 18:22:00 **网络出版地址**:http://www.cnki.net/kcms/detail/51.1596.T.20140710.1822.001.html

成功地解决了全局分析方法中计算量大的问题,并解决了全局分析中的不足,但却无法解决对于初检文档的依赖问题。

虽然,传统的查询扩展方法在技术上有了很大改进,但没有显著提高信息检索效率。近年来,概念语义相似度查询扩展逐渐成为该领域的研究热点,目前主要有基于本体和基于语义词典的查询扩展方法。在基于语义词典的方法中,利用知网的一系列的"义原"对概念进行描述[5],通过计算2个义原在层次体系中的路径来求义原的相似度,从而对查询项进行扩展。本体方法一般是通过对本体中查询项的上下位概念进行扩展,提高检索有效性。通过比较,发现基于本体的方法具有更强的概念结构层次关系,目前几种基于本体的方法[6-8],没有对本体展开全面的利用,只利用了本体中的某些侧面,概念间的相似度的计算不够科学,所以概念的语义不能被完全反映出来。

综上所述,为了从大量短文本数据中发现更多的特定领域敏感信息,利用可拓方法建立了可拓知识库,提出一种概念间相似度的计算方法和查询扩展算法,该算法充分考虑概念间的可拓关联关系与结构化信息,使得查询词具有明确的可拓关联性,隐含的的相关特征词将会被提取,结合可拓推理方法可以解决一般短文本数据中特征不足的问题,可提高相关领域敏感信息的发现率,提高了检测效率。

1 可拓模型

可拓模型是建立在基元基础之上的,基元在可拓学中是描述事、物及关系的基本元。可拓学中利用基元来形式化地表达信息。基于基元可建立知识模型,实现对知识的形式化描述[9-10]。

基于可拓学中的基元理论,可以建立特定领域中的可拓知识模型。例如需要描述某些敏感短文本信息,可以建立以下的物元模型:

$$R = \begin{bmatrix} \text{某组织}, \text{组织者}, \text{李某} \\ \qquad\qquad\quad \text{于某} \\ \qquad\qquad\quad \text{叶某} \end{bmatrix}$$

其形式为:

$$\begin{bmatrix} N, c_1, v_1 \\ \qquad v_2 \\ \qquad v_3 \end{bmatrix} = (N, C, V)。$$

同理,根据事元的定义,可用事元表示以动词为模型名的模型,如下列模型:

$$I = \begin{bmatrix} \text{静坐}, \text{施动对象}, \text{学员} \\ \qquad\quad \text{支配对象}, \text{某地} \end{bmatrix}$$

其形式为:

$$\begin{bmatrix} d, b_1, u_1 \\ \quad b_2, u_2 \end{bmatrix} = (d, B, U)。$$

根据关系元的定义,表示关系的模型可以用关系元来描述,如以下模型:

$$Q = \begin{bmatrix} \text{隶属}, \text{前项}, \text{某媒体} \\ \qquad\quad \text{后项}, \text{某组织} \end{bmatrix}$$

其形式为:

$$\begin{bmatrix} O_r, c_{r1}, v_{r1} \\ \quad c_{r2}, v_{r2} \end{bmatrix} = (O_r, c_r, V_r)。$$

物元、事元、关系元统称为基元,可以形式化表达事、物和关系。通过物元建立对事物的形式化描述,通过事元表达事物之间的动作关系,而关系元则用来表达事物之间的联系,通过这3种基元模型就可以把相对独立的事物有机的联系在一起,建立具有结构化与联系性的表达形式。

2 可拓知识

可拓知识是一般知识概念的扩展,是在可拓学基础上的对知识的形式化描述[11]。

2.1 拓展式

可拓学的拓展原理的表达式,即拓展式(发散式、相关式、可扩式、蕴含式等)是可拓知识的基础知识。

发散式:

$$(N_1, c_1, v_1) -| (N_i, c_j, v_k) \tag{1}$$

相关式:

$$(N_1, c_1, v_1) \sim (N_i, c_j, v_k) \tag{2}$$

蕴含式:

$$(N_1, c_1, v_1) => (N_i, c_j, v_k) \tag{3}$$

可扩式:

$$(N_1, c_1, v_1) \oplus (N_i, c_j, v_k) \tag{4}$$

其中,i, j, k 均为正整数。

2.2 变换蕴含式

可拓学的传导原理表示为变换蕴含式,它是变化的知识:$(T_u u = u') => ({}_u T_v v = v')$,可简写为:$T_u => {}_u T_v$。

可拓学中的拓展式是可拓知识的基础知识;具有传导特点的变换蕴含式则是一种特别的变化知识。它们共同构成了可拓知识,可将可拓知识总结为:

可拓知识=拓展式(基础知识)$\oplus$变换蕴含式(变化知识)。通过发散、相关、蕴含传导及可扩性可建立相互的强关联性。蕴含式与传统产生式规则类似,拓展式是传统产生式扩展。变换蕴含式是具有动态特性的变化知识,是解决矛盾问题的关键方

法。

2.3　可拓知识的关联性

拓展式及变换蕴含式决定了基元之间的关联性。这种关联性将将作为一种标记存储在可拓知识库中,由相关领域专家来维护这个库,下面给出物元关联网的描述。

关联网定义:给定一个物元 R,与 R 有关的所有物元与它一起构成的物元网,称为 R 的相关网,用符号表示为如图 1 所示,图中,R_1、R_2、R_3、和 R_4 与 R 之间是一种可拓相关关系,通过扩展到可扩关系、蕴含关系与变换蕴含关系,可建立更复杂的可拓关联网。图 2 为基于可拓关联关系建立的关于某组织的基元化的相关知识,可看出其可拓关联性。

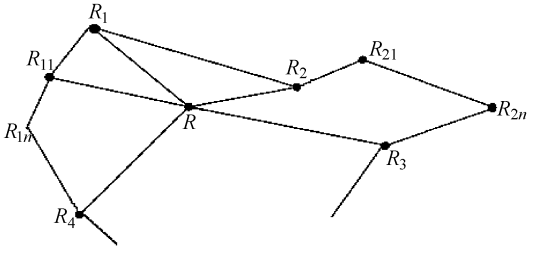

图 1　R 的关联网

Fig. 1　Correlation net of R

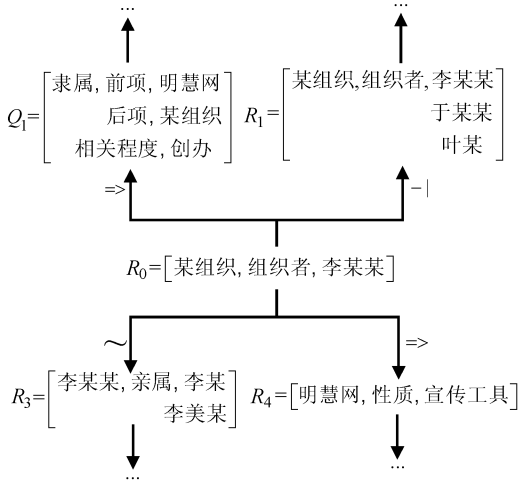

图 2　可拓知识关联性

Fig. 2　Correlation of extension knowledge

3　可拓知识库的构建

依据可拓学的基本理论,可拓知识库的

机理在于通过可拓变换将不相容问题转化为相容问题。可拓知识表示方法与其它知识表示方法在一定条件下可以互相转化。因此,可拓知识库与传统知识库是兼容的且是超集关系。

3.1　可拓知识库系统的组成

可拓知识库系统由主要由目标库、条件库、公用知识库、领域知识库、推理机组成。可拓知识库问题求解是以物元模型为基础 $P = G \times L$。其中,G、L 分别为目标与条件物元。多种可能的目标的集合即为目标库,其操作是每次仅选择一个目标。条件库主要存放条件物元及结果。公用知识库存放共用知识物元,领域知识库存放特定问题的知识物元。推理机是可拓推理技术的具体实现。

3.2　可拓推理与匹配

可拓推理是知识推理的扩展:

1)拓展推理

对拓展式的假言推理称为拓展推理。以物元的“一物一征多值”的发散式为例,发散式推理表示为:

$$(N_1,c_1,v_1) \wedge [(N_1,c_1,v_1) -| (N_1,c_1,v_i)] \vdash (N_1,c_1,v_i)。$$

2)传导推理

变换蕴含式是可拓变换与传导变换之间的蕴含式,传导推理表示为:

$$(T_u u = u') \wedge [(T_u u = u') => ({}_uT_v v = v')] \vdash ({}_uT_v v = v')。$$

3)菱形推理模式

定义:δ 代表某种或几种拓展推理方式,K 为关联函数约束,先拓展后收敛的菱形推理表示为:

$$(N_1,c_1,v_1) \xrightarrow[\delta]{} (N_1,c_i,v_i) \xrightarrow[K]{} (N_1',c_1',v_1')。$$

匹配:对于一个给定的上下文,可以利用其中的事实条件,通过相关网去匹配知识库中存储的知识物元。

3.3　问题求解策略

解决不相容问题必须通过可拓变换,其工具是关联函数,关联函数公式为:

$$k(x) = \rho(x,x_0,X_0)/D(x,X_0,X),$$

其中,$X_0 = \langle a,b\rangle$,$k(x) > 0$ 为正域区间。$X = \langle c,d\rangle$,$k(x) < 0$,是质变区间。关联函数本身属于知识。当 x 从区间 X_0 变化到区间 X 后,即关联函数 $k(x)$ 由正数变为负数,表明矛盾问题得到解决,问题求解流程如图 3 所示。

对于很多复杂的矛盾问题,直接变换可能无法使矛盾问题得到化解,此时可进行多种可拓变换包括传导变换的运算或复合,结合可拓推理,再求问题的关联度,并对规范后的关联度进行优度评价,从中选择优度值最高的即为问题解决方案,若在所有可能的变换与推理之后,优度值最高的方案依然不能使问题相容,则通过设定的中止参数退出循环,得出矛盾问题暂不能解决的结论。

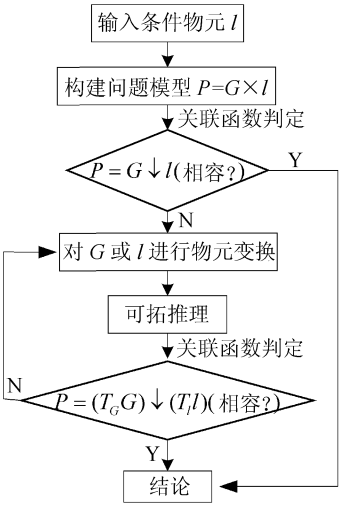

图3 问题求解流程

Fig. 3 Procedure of problem solving

4 基于可拓知识库的概念间相似度计算方法

在可拓知识库结构中,影响概念相似度的因素主要有概念间的关系权重、语义距离等。因此,在给出2个主要因素计算方法的基础上,利用可拓关联性来计算概念间的相似度。

4.1 关系权重

对于概念间的某种可拓关联关系,用关系权重w反映这种关系的关联程度,概念间的边将被赋予相应的权值。根据可拓学中的定义,各种关联关系的排序是:-| > ~ > => > ⊕ > T =>;且关联权重的范围在[0,1],其具体值的设置可根据情况而定,比如定义各类可拓关联关系的权重如表1所示。

表1 可拓关联及权重

Tab. 1 Extension correlation and weights

可拓关联类型	关联权重(w)
发散 -|	0.90
相关 ~	0.85
蕴含 =>	0.70
可扩⊕	0.70
变换蕴含 T =>	0.60

从表1中可以看出,关联权重的范围在[0,1],关联度越高,则相应权重会越大。

4.2 语义距离

将可拓知识库中的关联关系看成网,设其中2个概念为a和b,连接这2个节点的路径有$path(1)$,$path(2)$,…,$path(n)$,每一条路径经过m条边,每条边的关系权重为w_i,$i \in (1,n)$。

概念a和b在路径$path(i)$中的语义距离如式(5)所示:

$$Distance(path(i)) = \sum_{1}^{m} w_i/m \tag{5}$$

其中,m为路径$path(i)$经过的边数。

概念a和b的语义距离如式(6)所示:

$$Distance(a,b) = \mathrm{Min}(Distance(path(i))) \tag{6}$$

在语义距离基础上,定义概念相似度Sim。当a和b2个概念位于同一物元时,即两者为同一物元的2个不同属性值时,如式(7)所示,否则如式(8)所示:

$$Sim(a,b) = 1 \tag{7}$$

$$Sim(a,b) = \frac{\alpha}{Distance(a,b)} \tag{8}$$

其中,人工参数α的值由用来限定语义距离与概念相似度的关系,缓和路径长度的变化过快给相似度计算所带来的负面作用。

4.3 基于可拓知识库的转换

基于可拓知识库的概念查询扩展基本思想是:通过已经给出的的概念相似度的计算方法,量化概念间相似度结果并存储,查询检测时,查找与条件概念相似度大于阈值的概念,将其添加到到初始检测条件中,一并交给检索引擎。

初始输入检测条件$T = \{A,B,C\}$,经过可拓知识库扩展后的检测条件为$T' = \{A,B,C,A',B',C',\cdots\}$,$T \in T'$。

下面给出基于可拓知识库的短文本扩展算法:

```
1. T = {A, B, C}, extension knowledge base (EKB);
2. for( term in EKL)
   {
     Term -> term′;
     T′. add(term);
     if ( term in T′){
    d = dist(term_i, term_j);
    if(d > λ)//λ 为相似度阈值
     return d;
     S. add(term);
      // 将高于阈值的的概念添加到结果集}
   }
   return S; // 返回最终扩展集
```

本算法通过设置阈值λ提取出相似度高的几个概念添加到扩展集中,λ值可通过反复试验来平衡较高的准确率与召回率及实际运行效率的关系。

5　实验及结果分析

利用项目组收集的国内政治领域共 2 520 篇网民评论,设计 1 组对比实验,表 2 给出的是部分检索效果对比,下划线为算法检索出的部分,方法 1 与 2 的区别是后者增加了新的推理方法即菱形推理模式方法,表格显示所提出的新方法相比传统的检测方法对文本特定信息的检测更加全面。

表 2　实例效果对比

Tab. 2　Comparison of different methods

一般检测方法	提出的方法 1	提出的方法 2
做为极端重要的部门枢纽FBI内的基础设施中心持续运作发出针对计算机攻击的警告并根据策略做出反应	做为极端重要的部门枢纽FBI内的基础设施中心持续运作发出针对计算机攻击的警告并根据策略做出反应	做为极端重要的部门枢纽FBI内的基础设施中心持续运作发出针对计算机攻击的警告并根据策略做出反应
国家重要的系统和资产遭遇敌对国家的破坏将对国家安全经济安全产生巨大的灾难性影响	国家重要的系统和资产遭遇敌对国家的破坏将对国家安全经济安全产生巨大的灾难性影响	国家重要的系统和资产遭遇敌对国家的破坏将对国家安全经济安全产生巨大的灾难性影响
实施的此项计划是历来最为重视的其被用来设计网络空间保护方案的尝试活动	实施的此项计划是历来最为重视的其被用来设计网络空间保护方案的尝试活动	实施的此项计划是历来最为重视的其被用来设计网络空间保护方案的尝试活动
针对敏感部门政府会指派一名高级官员处理与私营企业的沟通协调其决定政企合作应对网络威胁与处理应急突发事件	针对敏感部门政府会指派一名高级官员处理与私营企业的沟通协调其决定政企合作应对网络威胁与处理应急突发事件	针对敏感部门政府会指派一名高级官员处理与私营企业的沟通协调其决定政企合作应对网络威胁与处理应急突发事件

表 3 给出的是针对 3 大主题的检测条件 A、B、C,根据通用的评价标准对比 3 种方式的综合检测性能对比统计结果。其中,方式Ⅰ为一般的短文本检索方法;方式Ⅱ为本文提出的基于可拓知识库扩展算法;方式Ⅲ为方式Ⅱ基础上增加菱形推理模式。最后使用准确率与召回率来衡量实验的效果:

Precision =

被检出相关文档量 / 被检出文档总量。

Recall =

被检出相关文档量 / 总文档中所有相关文档量。

实验采用了 3 组大类的检测条件进行测试,表 3 给出了实验结果对比。

表 3　实验结果对比

Tab. 3　Comparison of experimental results

检测分类	准确率 P/%			召回率 R/%		
	Ⅰ	Ⅱ	Ⅲ	Ⅰ	Ⅱ	Ⅲ
A	51.68	76.87	76.14	72.29	68.46	70.05
B	61.26	52.67	50.75	46.76	55.69	57.84
C	52.34	67.31	66.24	81.40	83.23	85.67

表 3 显示提出的新方法较方式Ⅰ在准确率、召回率 2 个指标上都有所提高,且由于增加了新的推理方式,方式Ⅲ的准确率虽不及方式Ⅱ,但在召回率上,前者效果显著。新方法从语义方面考虑了词之间的可拓关联性,借助于领域专家维护的可拓知识库,事实上随检测条件的逐渐增多,检测意图表达充分,所以能取得在 2 个指标上的优势,而方式Ⅰ只是词匹配概率的一般统计。

实验发现,提出的方法在检测时有着更高的准确率和召回率。但也存在缺点,即在扩展过程中需要查询项与可拓知识库中相关项的相似度的计算,增加了一定的时间开销。但随着计算机计算能力的增强,这种代价其实对检测效果的影响较小。

6　结束语

对传统的敏感信息查询扩展方法进行分析,提出了基于可拓知识库的查询扩展算法,并对算法的可行性进行了验证。实验证明借助于专家构建的针对某些敏感信息的可拓知识库的查询扩展算法能有效把握短文档中的关联关系,相比传统方法,能更好地解决短文本中敏感信息检测的问题,敏感信息的检测准确率、召回率都有一定程度的提高。

参考文献:

[1] Leung K W T, Ng W, Lee D L. Personalized concept-based clustering of search engine queries [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(11):1505 - 1518.

[2] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval[J]. ACM Computing Sur-

veys(CSUR),2012,44(1):1 – 56.

[3] Wan Jing, Wang Wencong, Yi Junkai, et al. Query expansion approach based on ontology and local context analysis[J]. Research Journal of Applied Sciences, Engineering and Technology, 2012, 4(16): 2839 – 2843.

[4] Lei Jiayin, Li Weijiang, Wang Feng, et al. A survey on query expansion based on local analysis[C]//Proceedings of the 4th International Conference on Intelligent Networks and Intelligent Systems. Kunming: IEEE Computer Society Conference Publishing Services, 2011.

[5] Jiang Min, Xiao Shibin, Shi Shuicai, et al. An improved word similarity computing method based on HowNet[J]. Journal of Chinese Information Processing, 2008, 22(5): 84 – 89. [江敏, 肖诗斌, 施水才, 等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报, 2008, 22(5): 84 – 89.]

[6] Segura A, Salvador-Sanchez, García-Barriocanal E, et al. An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene-ontology[J]. Knowledge-Based Systems, 2011, 24(1): 119 – 133.

[7] Batet M, Sanchez D, Valls A, et al. An ontology-based measure to compute semantic similarity in biomedicine [J]. Journal of Biomedical Informatics, 2011, 44(1): 118 – 125.

[8] Lee J, Min J, Chung C. An effective semantic search technique using ontology[C]//Proceedings of the 18th International Conference on World Wide Web. Madrid: Springer-Verlag, 2009: 1057 – 1058.

[9] Cai Wen. Extension theory and its application[J]. Chinese Science Bulletin, 1999, 44(17): 1538 – 1548.

[10] Men Baohui, Wang Zhiliang, Liang Chuan, et al. Application of matter element model to evaluating on of resources carrying capacity of regional groundwater[J]. Journal of Sichuan University: Engineering Science Edition, 2003, 35(1): 34 – 37. [门宝辉, 王志良, 梁川, 等. 物元模型在区域地下水资源承载力综合评价中的应用[J]. 四川大学学报: 工程科学版, 2003, 35(1): 34 – 37.]

[11] Chen Wenwei, Huang Jincai, Yang Chunyan, et al. extension knowledge and extension knowledge reasoning[J]. Journal of Harbin Institute of Technology, 2006, 38(7): 1094 – 1096. [陈文伟, 黄金才, 杨春燕, 等. 可拓知识与可拓知识推理[J]. 哈尔滨工业大学学报, 2006, 38(7): 1094 – 1096.]

(编辑 杨 蓓)