

倾向性文本的概念图过滤技术的研究

刘培奇, 凡 星, 段中兴

(西安建筑科技大学 信息与控制工程学院, 陕西 西安 710055)

摘 要: 本文提出一种基于概念图和简单语义分析相结合的文本过滤技术. 利用概念图知识表示方法定义文本中概念间的语义倾向关系, 把文本过滤问题转换成基于概念图的语义匹配问题. 在使用概念图匹配推理中, 利用敏感词汇分类和词语的褒贬性, 有效简化过滤的匹配处理. 经实际应用表明, 基于概念图的倾向性文本过滤方法可以有有效的识别和过滤倾向性文本信息.

关键词: 文本过滤; 概念图; 倾向性文本; 语义分析

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2012)12-0084-04

The Research of Conceptual Graphs Filtering Technology for Tendency Text

LIU Pei-qi, FAN Xing, DUAN Zhong-xing

(School of Information and Control Engineering, Xi'an University of Arch. and Tech., Xi'an 710055, China)

Abstract: This paper proposes a novel method of combination of conceptual graphs and simple semantic analysis, to solve the problem of low recall and precision rates. Using the formal of conceptual graphs to define the semantic orientation relationships of concepts in the text, and convert text filtering problem into conceptual graphs-based semantic matching problem. In order to simplify the process, reasoning matches with conceptual graphs on the basis of sensitive classification and words of part-of-speech. Practical application shows that the method can effectively identify and filter tendentiousness text message.

Key words: text filtering; conceptual graphs; tendency text; semantic analysis

1 引言

目前,对不良信息过滤主要采用关键词匹配技术,根据信息的主题内容进行过滤,可有效拦截敏感主题词汇.但对带有倾向性表述进行过滤时,存在误判的局限性.

目前,国内外对倾向性文本过滤问题进行了大量研究,主要研究方法有基于语义分析和基于机器学习两大类,并取得一定的研究成果^[1-2].但是由于自然语言自身表达的复杂性和灵活性,大多研究方法有待进一步改进提高.本文以自然语言理解理论为基础,在对主题词相关的主题信息进行分析的基础上,提出基于概念图知识表示的文本过滤系统^[3].

2 概念图基本理论

概念图(Conceptual Graphs, CGs)是计算机专家 John F. Sowa 提出的一种形式化知识表示工具,由概念结点、关系节点和有向弧组成的有穷连通图.可形式化定义为:

$$CGs = (\text{Concept}, \text{Relation}, F)$$

其中, $\text{Concept} = \{C_1, C_2, \dots, C_m\}$ 为 CGs 的概念结点集合,用来表示实体、动作、事件和状态等; $\text{Relation} = \{R_1, R_2, \dots, R_n\}$ 为 CGs 的关系节点集合,用来表示概念间的相互关系; $F \subseteq (\text{Concept} \times \text{Relation}) \cup (\text{Relation} \times \text{Concept})$ 为连接概念结点和关系节点有向弧的集合,表示节点间的作用关系.

收稿日期: 2012-04-19; 修回日期: 2012-05-18

基金项目: 国家自然科学基金项目(51178373)

CGs有图形表示和线性表示两种表示方法.其中,图形表示法用矩形框表示概念结点,用椭圆表示关系节点,用有向弧表示节点间的联系.例如,语句“we criticize falungong.”的CGs图形表示如图1.

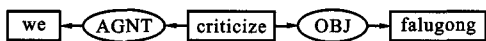


图1 概念图的图形表示法

图1中,“we”和“falun gong”都是一个实体,用一个概念结点表示,同样“criticize”作为一个动作,也通过概念结点表示;关系“AGNT”和“OBJ”就分别表示动作发起的“主体”和“受体”关系.虽然CGs的图形表示法形象直观,但为了便于在计算机内部表示和处理,还采用线性表示法,其中方括号表示概念结点,圆括号表示关系节点.图1的线性表示为

[critic]→(AGNT)→[we]
(OBJ)→[falugong]

概念图是一种强大的知识表示工具,不仅可以表示一阶谓词逻辑,还可以表示高阶逻辑和模态逻辑,具有丰富的推理功能,通过语言或语境的理解建立起具有严格的概念图的形式,可以有效避免歧义.

3 过滤系统结构模型

基于如下基本思想:

(1)首先用户向过滤系统提交待过滤文本后,进行有关主题词的匹配定位,考虑到有可能存在干扰词的情况,进行必要的扩展查询.

(2)其次由查询的结果确定主题对象,依据主题对象相关信息生成概念图.

(3)再次根据规则库中规则,进行基于概念图的匹配推理,确定每个主题对象所在句段的合法性.

(4)最后计算整个文档的相关度,最后根据设定的阈值得到文本的过滤结果.

本文设计的基于概念图的倾向性文本过滤系统模型如图2所示.

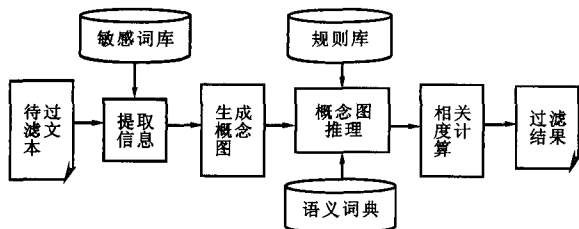


图2 基于概念图的过滤系统模型

在图2中,敏感词库用于存储相关敏感关键字,用于构造成待推理分析的主体,在本文中称之为主题对象.

4 系统过滤功能的实现

4.1 提取主题信息

由于自然语言具有极为复杂灵活的结构和表达,以当前的计算机认知能力,语义分析只能使用复杂的语义词典和语义框架来近似表达.考虑到对倾向性文本过滤的特殊性,词库中的主题对象不仅标注了基本的词性、词义等,尤其重要的是标注其本身的倾向性,可对倾向性进行量化.量化值用CHRC表示:(1)具有正面肯定倾向性量化值为1;(2)具有负面否定倾向性量化值为-1;(3)不具有倾向性不用赋予量值.

在文本中查找主题词时,由于自然语言表达形式的多样化和一些为躲避审查故意人为加入干扰字符的情况,可以通过主题相似度进行匹配计算^[4],还原变形的敏感主题字串.设有两个长度为 m 和 n 的主题字串分别用 $T_a=\{a_0, a_1, \dots, a_{m-1}\}$ 和 $T_b=\{b_0, b_1, \dots, b_{n-1}\}$ 表示.则主题字串 T_a 和 T_b 的相似度取决于两个子串中相同文字的位置,定义:

$$L(T_a, i, T_b) = \{k \mid T_b(k) = T_a(i), k = 0, 1, \dots, m-1\}$$

为 T_a 中的第 i 个字符与 T_b 中相匹配的字符的位置集合.当在 T_b 中没有出现与 $T_a(i)$ 相匹配的字符时,则 $L(T_a, i, T_b)$ 为空集.同时,定义:

$$D(T_a, i, T_b) = \begin{cases} n & L(T_a, i, T_b) = \emptyset \\ \min\{|k-1| \mid k \in (T_a, i, T_b)\} & \text{其他} \end{cases} \quad (1)$$

为文字 $T_a(i)$ 的最小匹配偏移值,定义:

$$CC(T_a, i, T_b) = \frac{n - D(T_a, i, T_b)}{n} \quad (2)$$

为文字 $T_a(i)$ 的匹配贡献值,则主题字串 T_a 和 T_b 的相似度为

$$\text{Sim}(T_a, T_b) = \frac{1}{m} \sum_{i=1}^m CC(T_a, i, T_b) \quad (3)$$

可见文本 T_a 和 T_b 的相似度值在 $[0, 1]$ 之间.一般根据不同的主题信息和过滤要求可自行设定相似度阈值.考虑到为了扩大匹配范围和存在二级过滤,相似度阈值不宜过大.当在确定某主题词后,就可以针对这个特定主题词进行倾向性判断过滤.

4.2 改进的概念图知识表示

传统的CGs知识表示中,概念结点并没有反映出概念自身的立场倾向性,更没有指出两个相连概念结点倾向性间的关系.鉴于此,本文在传统CGs理论上,把词语的褒贬倾向性引入到CGs知识

表示中,并通过研究发现关系节点两端的概念结点倾向性间的关系,将改进的概念图知识表示方法应用到倾向性文本过滤系统中。

在对包含主题对象的句子进行倾向性判断时,忽略材料、地点和类型等要素,因为这些要素并不能表示出句子所表达的立场,对文本的倾向性判定影响很小,而作为句子谓语的动词,却能表现出动作发出主题对动作受体的作用和立场.在概念图的生成过程中,当作为谓语的动词或动作主题、受体本身具有立场倾向性时,对应的概念图概念节点也要生成代表其倾向性的量化值 CHRC 节点.概念图部分关系的倾向性 CHRC 值生成情况如表 1 所示。

表 1 关系的倾向性标注生成情况表

CGs	CGs	生成	CGs	CGs	生成
关系	名称	情况	关系	名称	情况
AGNT	动作	生成	TYPE	类型	不生成
OBJ	对象	生成	MATR	材料	不生成
MANR	方式	生成	INST	工具	不生成
LOC	地点	不生成	DEST	目的	不生成

这样包含主题对象句段的局部倾向性 R 就可以按公式(4)得到

$$R = \prod_{i=1}^n CHRC_i \quad (4)$$

其中, $CHRC_i (i=1,2,\dots,n)$ 为生成概念图中每个表明节点倾向性的 CHRC 值.例如,依据 PLN-LP 文法^[5],句子 “We fight heresy through legal means.” 所对应的语法树如表 2 所示。

表 2 对应 PLNLP 语法树

DECL	NP	NOUN	“We”
	Verb	“fight”	
		NOUN	“heresy”
	Pp	prep	“through”
		NOUN	“legal means”

其中,DECL 表示这是一个陈述句,fight 表示它是一般现在时态,We、heresy、means 是名词形式,legal 修饰 means 构成短语 legal means,用来表现 fight 的方式,以及一些语法和词法特征.从这个语法树开始对这个句子进行语义解释,首先找出这个语法树的每一个叶子节点的正则图.因为,“We”是“Person”的所指,“We”对应概念 [Person: We],对应的正则图为

$$[Action-ENTITY] \leftarrow (AGNT) \leftarrow [fight] \rightarrow (OBJ) \rightarrow [heresy] \\ \downarrow \\ (MANR) \rightarrow [legal\ means]$$

由于动作受体“heresy”和谓语动词“fight”本身都具有否定倾向,它们在词库中对代表倾向性的 CHR 值都应为 -1;同时,具有肯定倾向 legal means 作为对 fight 的肯定,说明作为动作主题的 “we” 对一个具有否定性质的事物持否定立场,用公式(4)计算就是 $(-1) * 1 * (-1)$,结果 1 表示整个句子的立场是正倾向性的,实现现实意义.结果就得到整个句子的概念图,如图 3 所示。

$$[fight] \rightarrow (OBJ) \rightarrow [heresy] \rightarrow (CHRC) \rightarrow [-1] \\ (AGNT) \rightarrow [Person: We] \\ (MANR) \rightarrow [legal\ means] \rightarrow (CHRC) \rightarrow [1] \\ (CHRC) \rightarrow [-1]$$

图 3 整个句子的概念图

对出现敏感主题词的句段,应用文法规则,产生对应的语法树,最终可转化为概念图的知识表示形式。

4.3 基于概念图的推理

由于本系统中针对主题信息都是转化为概念图表示的,规则库中相应的规则也是通过概念图来表达的,所以采用基于概念图的正向推理策略.基于概念图的推理是通过概念图上的匹配为基础的,两个概念图的匹配包括完全匹配、投影匹配和最大连接匹配^[6].事实与规则的匹配过程,实际上也就是事实概念图与规则概念图的匹配过程.基于概念图匹配原理,系统过滤功能实现算法如下:

(1)由提取的主题信息生成事实概念图。

(2)设 u 是事实概念图, v 是规则前提中的一个图.如果 $u=v$,即 u 和 v 完全匹配,则匹配成功,转入(5);否则,转入(3)。

(3)如果 $u \leq v$,则调用投影运算过程,求出 v 在 u 上的投影,以及合一参数集合.若匹配成功,转入(5);否则,转入(4)。

(4)调用最大连接运算过程.如果 u 和 v 的最大连接存在,则求出此最大连接,以及合一参数集合,匹配成功,转入(5);反之,匹配失败。

(5)由匹配结果概念图计算出句段的局部倾向性 R 值,最终得出句段倾向性结果。

在对倾向性文本过滤时,对文本中出现词库中的主题对象,依据词典中主题对象对应的倾向值和规则库中的推理规则进行推理匹配,例如:在与邪教“法轮功”作斗争的过程中,我们应该坚持依法办事.设主题词“法轮功”是词库中已存在可用的主题对象,由于该主题词在词库中的倾向值具有否定倾向,所以赋予“法轮功”的倾向性的 CHRC 值为 -1. 分

析可知动词“fight”具有反对否定的倾向,所以在词库中对应的为否定值,转化为概念图中的倾向性 CHRC 值就为 -1,可得到对应的事实概念图 u 如下:

```
[Person; we]←(AGNT)←[fight]→(OBJ)→[falungong]→(CHRC)→[-1]
      →(MANR)→[lawful]→(CHRC)→[1]
      →(CHRC)→[-1]
      →(type)→[heresy]
```

当有规则 R 的概念图 v :

```
[person; x]←(AGNT)←[fight]→(OBJ)→[heresy]→(CHRC)→[-1]
      →(CHRC)→[-1]
```

依据匹配算法,则事实概念图 u 和规则概念图 v 的匹配结果如下:

```
[Person; we]←(AGNT)←[fight]→(OBJ)→[heresy]→(CHRC)→[-1]
      →(MANR)→[lawful]→(CHRC)→[1]
      →(CHRC)→[-1]
      →(NAME)→[falungong]
```

在以上的匹配过程中,由于关系节点 TYPE 连接“falungong”和“heresy”的两个概念结点,表示“falungong”作为“heresy”类型,前者与后者肯定是具有相同的倾向性的,并不需要对“heresy”进行重新赋予表示倾向性的 CHRC 值.而在规则 R 中,OBJ 关系节点连接的“heresy”作为动作“受体”,通过动作行为表示主题的立场倾向性,是需要赋予表明立场倾向性的 CHRC 值.在规则概念图 v 中,虽然出现带有否定倾向的敏感主题对象“heresy”,但是主体“person”的行为动作“fight”对主题对象具有否定倾向,同时在匹配结果中具有肯定倾向性“lawful”再次对主体加以肯定,表明整个事实句子表达的立场是倾向于合法的.

4.4 文本相关度

在对出现敏感词的句段进行概念图分析后,得出的只是此句段表达的局部倾向性,考虑到一个文本可能是包含正面信息的负面文本,也可能是包含负面信息的正面文本,所以单个语句的倾向性并不能代表整个文本的倾向立场.这就需要由每个包含主题对象的句段语义分析后得出的结果量化计算出全文的语义相关度^[7].设待过滤文本 D 的全文相关度 $R(D)$,则

$$R(D) = \sum_{i=1}^n Ri \quad (5)$$

其中, $Ri(i=1, 2, \dots, n)$ 是每个包含主题对象句段的局部倾向性.

通过全文的相关度与阈值进行比较,得出文本过滤结果.对不良信息的过滤可以依据不同的分级

设置不同的阈值,从而满足分类处理和不同需求.

5 实验结果及结束语

为检验本方法在实际系统中的应用效果,通过一些门户网站收集并筛选出主题对象的正相关、负相关以及无相关文本样本 300 篇(段),每类各 100 篇.在 Pentium 双核 E5300 环境下对系统进行测试,并使用基于统计方法的 kNN 分类方法进行了同样的测试.两种方法的测试结果表明:CGs 的召回率和准确率均好于 kNN.

采用本方法的系统对倾向性文本可以高效的进行识别和过滤,具有较高的召回率和准确率,进一步完善后,可应用于对倾向性文本的过滤.同时,本系统首次将概念图知识应用于倾向性文本过滤系统中,对其他领域的过滤系统的研究也具有一定的参考价值.

参考文献:

- [1] 朱嫣岚,闵锦,周雅倩,等.基于 HowNet 的词汇语义倾向计算[J].中文信息学报,2006,20(1):14-20.
- [2] YI J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques[C]//Proc of the 3rd IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2003:427-434.
- [3] Turney Peter. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4):315-346.
- [4] 王莹莹.中文短语相似度计算方法研究及应用[D].长沙:长沙理工大学,2008.
- [5] Stephen D Richardson, George E Heidorn, Karen Jensen. Natural language processing: the PLNLP approach[M]. New York: Kluwer Academic, 1993.
- [6] SOWA JF. Conceptual structure[M]. New Jersey: AddisonWelsley, 1984.
- [7] 吕滨,雷国华,于燕飞.基于语义分析的网络不良信息过滤系统研究[J].计算机应用与软件,2010,27(2):283-285.

作者简介:

刘培奇 男,(1959—),博士,副教授.研究方向为机器学习和数据挖掘.

凡 星 男,(1986—),硕士研究生.研究方向为数据挖掘、自然语言处理.

段中兴 男,(1969—),博士,教授.研究方向为智能信息处理与智能控制、智能检测、信息融合.