基于本体的话题检测与跟踪技术

刘 炜[1],李 明[1],杨合立[2]
(1. 兰州理工大学 计算机与通信学院,甘肃 兰州 730050; 2. 兰州理工大学 教务处,甘肃 兰州 730000)

摘 要: 基于前人在 TDT 中对语义矢量的相似性计算研究,以及本体和语法结构在文本相似性研究方面的应用成果,提出了以词频分析作为辅助手段,将新闻中的关键要素归纳为时间、空间、参与事件的主客体、行为等几个语义类; 借助 WordNet 与本体技术计算文档特征词的相似度,并且结合文本的语法结构特点,共同应用于文本的相似度计算,并以此作为新事件检测中相似度计算的基础,提高新事件检测的准确性。

关键词: TDT; 本体; WordNet; 文本相似度; 新事件检测; 语义矢量
中图分类号: TP391. 1

话题检测与跟踪(Topic Detection and Tracking, TDT) 的概念最早产生于 1996 年,当时美国国防高级研究计划署(DARPA) 根据自己的需求,提出要开发一种新技术,能在没有人工干预的情况下自动判断新闻数据流的主题。这项技术旨在帮助人们应对日益严重的互联网信息爆炸问题,对新闻媒体信息流进行新话题的自动识别和已知话题的持续跟踪。TDT 的研究方向主要分为 5 个组成部分,即报道切分、报道关联性检测、话题检测与跟踪以及针对各项任务的跨语言技术。TDT 研究的 5 个部分中,话题的关联性检测是所有任务中最主要的研究项目,只有提高对报道相似性判断的准确性,才能有效地检测和跟综后续的话题。本文研究的重点即是通过引入本体,提高对报道文本特征的概念相似度的计算,准确判断话题的相关性。

关联性检测的主要任务是检测随机选择的两篇报道是否论述同一话题。传统基于概率统计的 TDT 研究,报道与话题或者报道与报道之间的相关性,都是通过检验两者之间共有特征的覆盖比例进行评判。大部分针对关联性检测的研究都将问题的重心集中于文本描述以及特征选择。James Allan [1] 是最早使用自然语言处理技术(NLP) 解决 TDT 问题的学者之一。其采用 VSM 描述话题和报道,并对模型中的命名实体赋予更高的权重,以此执行 TDT 中的新事件检测(NED) 任务。Nallapati [2] 对这种方法进行了改进,其首先将特征划分到不同的语法类别,比如词性中的名词类和动词类,以及命名实体中的时间类、地点类和人名类。在这个基础上采用语言模型的概率统计方法,估计特征产生于不同语法类别的概率,并以此标记特征的权重。Juha Makkonen [3] 提出构建语义类,通过语义类的相似度来判别文档相似性的方法,以解决 TDT 中的关联性检测问题。

WordNet 是 Princeton 大学的一组心理词汇学家和语言学家于 1985 年开始开发的一部语义词典数据库,WordNet 根据词义而不是词形来组织词汇信息。WordNet 包含名词、动词、形容词和副词。WordNet 按语义关系组织词典。由于语义关系是一种词义之间的关系,而词义可以用同义词集合来表示,因此很自然的把语义关系看作为同义词集合之间的一些指针。语义关系有同义关系、反义关系、上下位关系和部分关系等 [4] 。从 WordNet 的这些特点可以看出,WordNet 与本体非常相似。TDT 是针对自然语言报道而进行的研究,在文本的相似性检测方面,对于地理名词的相似性,可以通过地理本体来扩展其语义特征。对于报道中的名实体和行为概念,则可以通过 WordNet 来计算其概念信息度。

以往多数的文档相似性研究对句子的结构及句子中的词性未作深入的分析,检索过程中忽略了文本的语法结构信息。S. Shehata 等 [5] 分析了句子的动词结构,结合传统的 TF - IDF 技术,将文档句子结构中不同的、重要的概念赋以不同的权重,改进了 TF - IDF 技术中出现次数相同的词有相同的权重的缺点,但该文献在句子结构的分析上还不够深入。黄承慧等人 [6] 提出了基于主谓宾结构的文本语义检索算法,该算法充分利用了句子的结构特征,在计算句子的相似度研究中取得了比较好的效果。

基于前人研究成果,提出对文档特征矢量选取的看法,并在 TDT 研究中通过对本体技术的引入,拓展了文档概念相似度技术在 TDT 中的应用,并借助句子结构信息提高文档相似度的计算精确性。在

特征词的使用方面,给特征词增加重要度的方法也是一种新的提法。通过以上技术,构造出新的 TDT 系统框架。

1　语义类模型

1.1　语义类分析

TDT 系统模型主要有概率模型和向量空间模型,向量空间模型不断发展,到后来由几个向量来表示一篇文档,进而有将文档用语义类[7]来表示文档的方法。Juha Makkonen [8]的研究中,一篇文档划分为 4 个语义类: LOCATIONS、TEMPORALS、NAMES 和 TERMS 类。这 4 个语义类的划分是通过事实观察得出的: 某个事件是发生在特定某个时间段,特定地点,有某些特殊的人或事物相关。这些特征通过整理得出这 4 个分类。这种划分方式仍然不够细致。通过分析,事件不但和某些特殊的人或事物相关,而且还会涉及一些特殊的行为。一个事件,是某些行为主体在某个时空范围内发生某些行为造成的,涉及到 4 类特征: 时间、地点、与事件相关的人或事物、所产生的行为。分析认为,这 4 类特征必不可少而且唯一的决定一个事件,而且这 4 类特征是一个事件必不可少的元素,所以,将这 4 类特征作为语义类是比较合理的划分方法。本文中,特征词被划分为 TEMPORALS、LOCATIONS、ENTITIES、ACTIVITIES 等 4 个语义类。

语义类这样划分还有一个优点,就是在句子结构分析时会涉及到主、谓语成份,这个在后文中可以看到,句子三元组的构成需要一些名词性和动词性的特征词,而这些特征词都存在于 ENTITIES 和 ACTIVITIES 类。

1.2　词项处理

一篇文档首先要经过词语切分和识别提取出其中的特征词。这里特别指出,由于汉语文本中的歧义较多,汉语语句的切分有一定难度,国内在文本消歧方面也有很多的研究。文档经过词语切分以后,就要进一步识别出词段的词性,这就需要查找语义词典给相应的词段赋于词性,对于多词性的词段要通过所在句子的句法构成来进行词性判别,找到符合句法逻辑关系的词性,这时候也要用到一个句法判别系统。在判定词性的过程中,要将一些对文章特征贡献不大的词舍弃,留下对文档比较有用的特征词,比如描述时间和地理位置的词段,及名词性词段、人名、专有名词和动词词段。这些词将按照前面所分的语义类进行归类并计算其对文章的重要度。

对于描述时间段的词段,在预处理阶段要尽可能的规范化,即将其映射在全局统一时间轴线上。如果文中有某个明确的时间点能使某个时间描述映射在统一时间轴线上,则这个时间描述应该保留,否则就应该舍弃; 因为无法确定的时间描述是无法计算其相似度的。

整个过程形式化说明如下。

定义 $D=\{d_1,d_2,\cdots,d_{|D|}\}$,$D$ 表示一篇文档。

筛选 D 中的特征词并划分为 4 个语义类来表示,即 $D=\{T,L,S,A\}$。

其中 $T=\{t_1,t_2,\cdots,t_{|T|}\}$,$L=\{l_1,l_2,\cdots,l_{|L|}\}$,$E=\{e_1,e_2,\cdots,e_{|E|}\}$,$A=\{a_1,a_2,\cdots,a_{|A|}\}$ 分别表示 4 个语义类 TEMPORALS、LOCATIONS、ENTITIES 和 ACTIVITIES。E 中的词为名词词性,A 中的词为动词词性。

语义类中的特征词都要增加重要度属性,以 T 为例,T 变为 $\{(t_1,\mathrm{IMP}_{t1}),(t_2,\mathrm{IMP}_{t2}),\cdots,(t_{|T|},\mathrm{IMP}t_{|T|})\}$,其他语义类与此类似。

2　相似度计算

2.1　重要度值的确定

由于新闻报道都有一个主题,有些词段可能和这个主题非常接近,而有些词段可能与主题关系不是很密切。在计算文档的相似度时,如果一个词段在文中出现的频度较高,则认为该词段更接近主题,对于文章相似性的比较来说更重要一些。词段的这个特性要通过重要度来体现,重要度的计算采用以下方法。

首先,计算名词和动词性特征词在文档中的词频。然后,将这些特征词按词频值的高低降序排列,排列后的自然数序号分配给该特征词作为其对该文档中的重要度,即频度越高的词其重要度的数值越小。对于 TEMPORLS 和 LOCATIONS 类的特征词,其重要度不能以其词频来决定,则将其与所在段落,所在句子中的名词和动词的特征词词频高的重要度对应。

2.2　TEMPORALS 类的相似度计算

设 T_x、T_y 为两个文档的 TEMPORALS 类,t_x、t_y 分别是它们的两个时间段分量,单位为天,$t_x\in T_x$,$t_y\in T_y$,则 t_x 和 t_y 的相似度为

$$\mathrm{SIM}(t_x,t_y)=2*\ \mathrm{INTERSECTION}(t_x,t_y)/(t_x+t_y)$$

式中: $\mathrm{INTERSECTION}(t_x,t_y)$ 表示 t_x、t_y 这两个时间段的交集,单位为天。

这里两个时间段的相似度要再用其重要度进行

修正。

$$SIM'(t_x,t_y) = 0.8SIM(t_x,t_y) + 0.2\max^{-1}(IMP_{tx},IMP_{ty})$$

式中: $\max^{-1}$ 表示最大值的倒数。

这里用重要度进行修正只是将重要度最大值取倒数并与前面计算得到的相似度值作了加权,因为重要度值越大表示该特征词的排序越靠后,对文档主题越不重要。即使是所比较的两个特征词词形完全一致,但因为其对文档的重要度不同,进行修正后相似度也不会是 1。由于计算中给重要度倒数值分配了 0.2 的权重,因此修正的结果也不会因为重要度值的介入而变化过大。

假设 T_x 中有 n 个时间段,T_y 中有 m 个时间段。T_x 中的某个时间段 t_x 要和 T_y 中的 m 个时间段都比较一次,就会有 m 个结果,选取其中的最大值表示 t_x 与 T_y 的相似度。当 t_x 中的 n 个时间段都和 T_y 中的 m 个时间段进行比较过后,就会有 n 个这样的最大值。计算这些最大值的算术平均值即为 T_x,T_y 的相似度,即

$$SIM(T_x,T_y) = \sum_{i=1}^{n}\max(SIM'(t_x,T_y))/n$$

2.3 LOCATIONS 类的相似度计算

地理位置的相关性最适合借助地理本体来计算,任何两个地理位置的相似度用其在地理本体中路径的相似性来表示。用这两个地理位置的公共路径的 2 倍去除以它们分别到根节点的路径总和,用它们的商来表示这两个地理名词的相似度[8]。

假设 L_x、L_y 为两个文档的 LOCATIONS 类,l_x、l_y 分别是 L_x、L_y 的两个分量,$l_x \in L_x$,$l_y \in L_y$,则 l_x 和 l_y 的相似度为

$$SIM(l_x,l_y) = 2*\lambda(LAST-SHARE(l_x,l_y))/\lambda(l_x)+\lambda(l_y))$$

式中: $\lambda(x)$ 表示在树状地理本体中地理名词节点到地理根节点的距离; $LAST-SHARE(l_x,l_y)$ 表示两个地理名词最后一个共享节点。

与 TEMPORALS 类类似,两个地理名词段的相似度要用重要度进行修正。

$$SIM'(l_x,l_y) = 0.8SIM(l_x,l_y) + 0.2\max^{-1}(IMP_{l1},IMP_{l2})$$

与 TEMPORALS 类类似,L_x,L_y 的相似度为

$$SIM(L_x,L_y) = \sum_{i=1}^{n}\max(SIM'(l_i,L_y))/n$$

2.4 ENTITIES 和 ACTIVITIES 类的相似度计算

ENTITIES 类和 ACTIVITIES 类的相似度计算采用基于 WordNet 的计算特征词语义距离的方法。

WordNet 是层次树结构。概念在 WordNet 树中有深度和宽度,深度是指概念离根节点的路径长度,宽度指概念孩子节点的个数。一般来说概念的深度越大,说明概念的分类越细,概念的相似度就越高;对于同一深度的概念,概念的宽度越大,表明概念划分的比较细,概念的相似度就高。因此,在计算两个特征词的语义距离时应该考虑其深度和宽度对相似度的影响,相应的修改语义距离的权重。

语义距离的计算如下[9]。

$$DIST(C_1,C_2) = \sum_{i=1}^{n} w_i$$

式中: C_1,C_2 表示两个概念; w_i 表示两个概念之间最短路径上的权重。

定义 $Dep(C)$ 为概念 C 在 WordNet 中的深度,$Dep(C)$ 表示概念 C 到根节点的路径长度。定义 $Wid(C)$ 为概念 C 在 WordNet 中的宽度,$Wid(C)$ 表示概念 C 的子节点的数目。定义 $Weight(C)$ 为由概念 C 引出的边的权重[9]。

$$Weight(C) = \begin{cases} \dfrac{1}{Wid(C)} \\ \dfrac{1}{Wid(C)} \times \dfrac{1}{2} \times Weight(parent(C)) \end{cases}$$

式中: 当 C 为根节点时,$Weight(C) = 1/wid(C)$; 当 C 为其他节点时,$Weight(C)$ 为其父节点的权重乘 1/2,再除以其宽度。这里如果 C 节点如为叶子节点,则令 $1/wid(C) = 1$。

这样当 C 为根节点时,深度为 0,权重为其宽度的倒数,宽度越大,权重越小,距离也就越短。当 C 为其他节点时,随着深度的加深,到节点 C,每次给上位节点的权重都乘 1/2 并除以其宽度,C 的权重就越来越小,其路径也就变短,表示概念 C 的权重随深度的加深和宽度的加大,距离也越短。

按照两个概念间的语义距离越短,两个概念越相似的原则,列出一个求两个概念相似度的公式[9]。

$$SIM(C_1,C_2) = 1 - \sqrt[t]{\frac{1}{2}DISC(C_1,C_2)}$$

当求得 ENTITIES 类或 ACTIVITIES 类中两个特征词的相似度后,语义类的相似度计算都类似之前 TEMPORALS 类的算法,再不赘述。

2.5 句子结构的相似度计算

采用句法分析工具 LingPipe[10],分析句子的主谓宾成份并构造句子的三元组形式。这个三元组形式将是计算句子结构相似度的依据。

在句子三元组的基础上增加了时间和地理的特

征词,变成句子五元组。句子中的特征词在比较计算时不需要用重要度做再修正,这是因为这里的计算是比较两个句子语义是否相似,不必考虑特征词是否接近文档主题。虽然这里不必考虑特征词的重要度,但由于句子对文章主题的重要性不同,因此也给句子三元组增加了重要度的属性。句子的重要度为 $IMP_s = \max(IMP_{ei}, IMP_{ai} \text{ in } ST)$,表示这个值为句子中所有名词和动词的重要度值的最大值。

定义句子 $ST = \{s, v, o, t, l\}$,s、v、o 分别为作主语、谓语、宾语的特征词集合,$s, o \subset E$,$v \subset A$;t、l 分别表示与句子相关的时间或地点的特征词集合 $t \subset T$,$l \subset L$。

假设 ST_1,ST_2 为两篇文档中的两个句子,s_1,s_2 分别为它们的主语。s_{1i},s_{2j} 分别是其主语中的特征词,特征词相似度的计算方法见前文。则 s_1,s_2 的相似度为

$$SIM(s_1, s_2) = \sum_{i=1}^{n} \max(SIM'(s_{1i}, s_2)) / n$$

其他成份的相似度计算与此类似。

计算出这些句子成分的相似度后,则两个句子的语义相似度为

$$SIM(ST_1, ST_2) = w_1 * SIM(s_1, s_2) + w_2 * SIM(v_1, v_2) + w_3 * SIM(o_1, o_2) + w_4 * SIM(t_1, t_2) + w_5 * SIM(l_1, l_2)$$

式中: w_i 为相应成份的权重,$i = 1, 2, 3, 4, 5$。

最后,通过句子的重要度对两个句子的相似度进行修正。

$$SIM'(ST_1, ST_2) = 0.8 SIM(ST_1, ST_2) + 0.2 \max^{-1}(IMP_{s1}, IMP_{s2})$$

2.6　整篇文档的相似度计算

当计算出了五个语义类矢量的相似度和句子结构的相似度后,就可以求出两篇文档的相似度。一般文档的相似度是这些相似度分量的一个加权和。

设 $SIM(T_1, T_2)$,$SIM(L_1, L_2)$,$SIM(E_1, E_2)$,$SIM(A_1, A_2)$ 分别是 TEMPORALS、LOCATIONS、ENTITIES 和 ACTIVITIES 的相似度,$SIM(ST_1, ST_2)$ 是句子结构的相似度,则两篇文档 D_1,D_2 的相似度为

$$SIM(D_1, D_2) = w_1 \times SIM(T_1, T_2) + w_2 \times SIM(L_1, L_2) + w_3 \times SIM(E_1, E_2) + w_4 \times SIM(A_1, A_2) + w_5 \times SIM(ST_1, ST_2)$$

这个公式中的权重 w_i 的值可以通过用典型数据对系统进行训练调整,确定出比较合理的 w_i 值。

3　结论

话题检测与跟踪技术自提出以来到现在技术不断深入,研究任务也不断提高。尽管如此,该项技术的核心研究即文档的关联性检测方面仍需要不断挖掘其潜力,以降低文档相似度比较中的漏检和误检。本体作为一种新出现的技术目前在话题检测与跟踪技术中的应用比较少,将本体应用于 TDT 研究也促进了 TDT 技术的发展。

在前人所提出的文档语义类矢量这种特征词的分类法的基础上进行改进,提出将动词特征词单列为一个语义类。在特征词的相似度计算方面,引入本体技术,通过语义词典 WordNet 计算特征词的概念相似度,改进系统只能识别同词形的词,而对同义词漏检的缺陷。提出给特征词增加重要度属性的办法,使文档的主要特征更突出。增加了句子结构相似度的判断,借此提高文档相似度判断的准确性。

话题检测和跟踪技术的研究,要由自然语义理解技术来推动,本体技术在自然语义理解方面有很大的潜力,可以相信将来本体技术在 TDT 研究中会得到更广泛的应用。

参考文献:

[1] ALLAN J, JIN H, RAJMAN M, et al. Topic – based novelty detection [C]//Proceedings of the Johns Hopkins Summer Workshop [C]. CLSP, Baltimore, 1999.

[2] NALLAPATI R. Semantic language models for topic detection and tracking [C]//Proceedings of HL T – NAACL2003 Student Research Workshop [C]. Edmonton, 2003.

[3] MAKKONEN J. Investigations on event evolution in TDT [C]//Proceedings of HLT – NAACL 2003 Student Research Workshop [C]. Edmonton, 2003: 43-48.

[4] MILLER G A, BECKWITH R, FELLBAUM C, et al. Introduction to WordNet: An on – Line lexical database, in Five Papers on WordNet, CSL report [R]. Princeton, USA: Princeton University, 1993.

[5] SHEHATA S, KARRAY F, KAMEL M. A Concept – based Model for Enhancing Text Categorization [C]// Proceedings of KDD 2007. San Jose, California, USA, 2007: 629-637.

[6] 黄承慧,印鉴,侯昉. 一种基于主谓宾结构的文本检索算法 [J]. 计算机科学,2010,37(9): 173-176.

[7] MAKKONEN J, AHONEN – MYKA H, SALMENKIVI M. Topic detection and tracking with spatio – temporal evidence [C]//ECIR, Pisa, 2003: 251-265.

[8] 李熙,徐德智. 基于 WordNet 的概念语义相似度研究 [J]. 湖南科技学院学报,2008,29(12): 115-116.