一种基于本体的文本主题提取方法研究

高玮军 马栋林 张其文
(兰州理工大学计算机与通信学院 甘肃 兰州 730050)

摘 要 在深入分析当前流行的文本主题提取技术和方法的基础上,提出一种将本体技术应用于文本主题提取的方法。使用本体技术用语义向量表示文本句,对文本进行预处理,然后进行语义相似度计算和语义聚类,最后从每类中抽取代表句生成文本主题。实验结果表明,该方法在提取文本主题方面是一个有效的方法。

关键词 文本主题 自动文摘 本体 文本聚类 语义向量

中图分类号 TP391 **文献标识码** A

AN ONTOLOGY-BASED TEXT TOPIC EXTRACTION METHOD RESEARCH

Gao Weijun Ma Donglin Zhang Qiwen
(*School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, Gansu, China*)

Abstract After thorough exploration with present text topic extraction technology and methods, the paper proposes an approach to apply ontology technology to text topic extraction. The approach counts on ontology technology to utilize semantic vectors to represent text sentences. Then it carries out semantic similarity calculation and semantic clustering. Finally it extracts representative sentences from every class to generate text topics. Experimental results show that the approach is feasible in text topic extraction.

Keywords Text topic Automatic text summarization Ontology Text clustering Semantic vector

0 引 言

文本主题提取的研究已经走过了半个世纪的发展历史。随着电子出版系统和国际互联网的快速发展,电子文献大量出现,文本自动主题提取的价值越来越大。1995 年,国际期刊 Information Processing & Management 出版题为 Summarizing Text 的专刊[1],标志着自动主题提取的时代已经到来。文本主题的自动提取是使用计算机从文本原文中选择最能代表文章主旨的词句,生成连贯、准确的主题描述。

主题自动提取是自然语言处理中的一个重要分支,在文本挖掘和信息检索领域中有着非常重要的用途。目前,主题提取算法可以分为三类:(1) 领域无关的方法[2],主要是利用统计技术,如向量空间模型、项频率、反转文档频率,同时利用语言学知识,如词汇相关性,来识别文档中的重要段落和语句。它的核心思想是:根据特殊的统计特征,计算每个语言单元(通常是句子)的重要度,最后将最重要的句子抽取出来,形成主题。这类方法的优点是与领域无关,缺点是主题的准确率低,连贯性差,如主要内容缺失,指代词悬挂,文摘句过长等。(2) 领域相关的方法,是基于语义分析和理解的主题提取方法,它需要较成熟的人工智能技术和大型的专家知识库,对文章进行深层的句法和语义分析[3-6]。典型的理解式方法使用预定制的模板,从原文中提炼重要的信息填入模板中,从而生成主题。它的主要缺点是,由于需要庞大的专家知识库和完善的语言学规则,只能应用到某些特定题材的文体和内容具有相当可预见性的文章中。(3) 将前两种方法融合得到的方法[7-10]。

随着人们对知识表示、信息组织和复用以及为用户提供有效服务的需求越来越强烈,本体(Ontology)作为一种能在语义和知识层次上描述信息系统的概念模型建模工具,自提出以来就引起了国内外众多科研人员的关注,并在计算机的许多领域得到了广泛应用。

从知识共享的角度看,本体可以被看作是一种概念化的显式说明,是对客观存在的概念和关系的描述。它起源于哲学概念,被哲学家用来描述事物的本质,后来被引入到人工智能的研究领域,在开发知识系统时用于领域知识的获取。Stilder 在进行深入研究后,认为本体是共享概念模型的明确的形式化说明,包括概念化、明确、形式化和共享四层含义。

本文在对目前流行的主题提取技术和方法进行深入分析和比较的基础上,提出了一种基于本体的主题提取方法。该方法将语义分析融入统计算法,同时结合语义聚类技术进行主题抽取。

本文提出的主题提取方法利用文本中句子之间的语义相关性,实现文档主题的自动生成。该方法结合了本体技术,包括语义标注、基于本体的语义特征向量、语义相似度以及基于本体的聚类算法等。该方法有以下特点:(1) 领域无关性,因此是通用的方法。(2) 利用抽取代表句的方法来产生主题文本。(3) 和文本聚类技术结合,使生成的主题文本有较好的可接受度。

收稿日期:2011-10-10。国家自然科学基金项目(61064011)。高玮军,讲师,主研领域:软件工程,模式识别。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1　基于本体的文本主题提取方法

1.1　文档预处理和语义标柱

为了用模型表达文档,需要对文档进行预处理,预处理中最重要的是分词。中文分词的难点是词的界限不清楚。文章是字的序列,词与词之间没有间隔标记,使得词的界定没有自然标准,而分词结果是否正确需要有一个通用、权威的分词标准来衡量。切词方法一般分为三种: 基于字符串匹配的切词方法、基于理解的切词方法和基于统计的切词方法。

为了提高效率,本文引入 N-gram 分词算法,并对它进行了改进。先建立一个停用词表,将停用词表中的词作为标点符号,与分割符一起,将原文档分隔成字串集合 $M1$,再将 $M1$ 分割成长度小于 3 的字串集合 $M1'$ 和长度大于等于 3 的字串集合 $M1''$。$M1'$ 不参加分词过程,直接作为待筛选的词,$M1''$ 参与分词,得到词语集合 M',合并 $M1'$ 和 M' 得到分词结果。之后还要滤去高频词、低频词和冗余词等,才能得到最终的分词结果 M。

文本分词完成之后,还要进行语义标注。将分词结果 M 中的词与本体概念逐一比对后,就可以将文档中的本体概念抽取出来。本文提出一种对单词和本体概念相似度进行计算的方法,设本体中有 N 个概念 $\{c_1, c_2, \cdots, c_N\}$, $c \in \{c_1, c_2, \cdots, c_N\}$,定义单词 W_k 与本体概念 c 的相似度为: $Sim_{kc} = P(W_k \mid c)$。$P(W_k \mid c)$ 表示单词 W_k 在 c 中出现的概率。

$$Sim_{kc} = P(W_k \mid c) = \begin{cases} \dfrac{N(W_k, c)}{W_c + \sum_i N(W_i, c)} & \text{if} \quad N(W_k, c) \neq 0 \\ \dfrac{W}{W_c + \sum_i N(W_i, c)} \times \dfrac{1}{W - W_c} & \text{if} \quad N(W_k, c) = 0 \end{cases} \tag{1}$$

其中, $N(W_k, c)$ 是训练集中单词 W_k 出现在包含概念 c 的文档中的次数, W_c 是概念 c 中的单词个数, W 是所有类中的单词个数,公式中使用了平滑方法以避免在训练集中没有出现过的单词概率取值为 0。

1.2　基于本体的语义特征向量

在经过文档预处理和语义标柱之后,对文档描述内容有了比较清晰的轮廓,就可以建立基于本体的语义特征向量。

针对传统向量空间模型在语义处理方面的缺陷,把本体引入向量空间模型,提出了一种基于本体的语义特征向量。对传统向量空间模型中语义项权重的设计、不同关键字之间的语义关系,以及语义特征向量的相似度计算等方面进行改进。

与传统向量空间模型不同,传统向量空间模型首先建立有单词或短语为关键词的字典,然后将每个文档表示成一个多维向量,最后用二进制向量或反转文档等方法将文档表示出来。在基于本体的语义特征向量中,使用本体库代替包含关键词的字典,利用文档描述的一个或多个概念及其属性所组成的语义向量来表示文档,从而完成语义化。

每篇文档都会描述特定的主题和内容,文档的每一部分也都会对某一概念对象进行描述,首先抽取出文档中的关键字,然后使用本体中的一个或多个概念及其属性来对其进行抽象和概括。对每个概念而言,从文档中抽取其特定属性及属性值,共同组成刻画文档内容的语义向量,如图 1 所示。

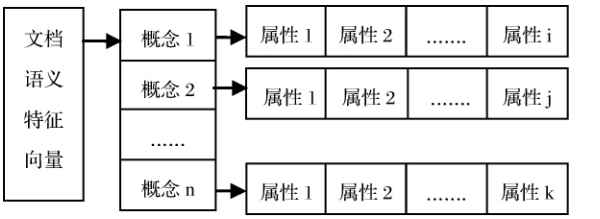

图 1　文档语义特征向量

1.3　基于本体的语义相似度分析

设整篇文章总共有 N 个句子,在传统向量空间模型中,可以得到 N 个向量:

$$v(d_i) = (w'_{i1}, w'_{i2}, \cdots, w'_{ik})$$

其中, $i = 1, 2, \cdots, N$。k 为文章中的单词个数, w'_{ij} 为第 j 个词在句子 i 中的权值。然后根据式(1)计算出第 j 个单词与本体概念 c 的语义相似度 Sim_{jc}。定义概念权重如下。

$$g_c = \sum_{k=1}^{W} Sim_{jc} \tag{2}$$

其中 g_c 是概念 c 的权重, W 是所有类中单词的个数。对 g_c 进行归一化,使其取值区间为 [0,1]。

如果存在这样一个本体概念 c,且 c 中的单词在 d_i 和 d_j 中共同出现,则概念 c 对 d_i 和 d_j 的相似度贡献就会比较大。鉴于此,定义语义向量 s_i 与 s_j 的语义相似度。

$$Sim(s_i, s_j) = \frac{1}{|s_i| \times |s_j|} \sum_{c=1}^{N} (g_c \cdot f) \tag{3}$$

其中, $|s_i|$ 表示语义向量 s_i 中的概念个数, g_c 是概念 c 的权重, f 是概念 c 中的单词在语义向量 s_i 与 s_j 中的共现频率。

根据式(3),对文档中的 N 个句子可以计算两两之间的相似度 s_{ij},最后的结果以上三角矩阵表示:

$$\begin{bmatrix} 0 & s_{12} & s_{13} & s_{14} & \cdots & s_{1N-1} & s_{1N} \\ & 0 & s_{23} & s_{24} & \cdots & s_{2N-1} & S_{2N} \\ & & 0 & & \cdots & \vdots & \vdots \\ & & & & \ddots & & \\ & & & & & 0 & S_{(N-1)N} \\ & & & & & & 0 \end{bmatrix}$$

存储矩阵所需的空间为 $\frac{N(N-1)}{2}$,计算矩阵的时间复杂度为 $O(N^2)$。

1.4　基于本体的聚类方法

基于本体的聚类算法如下:

(1) 对文章进行预处理和语义标注,获得文章中的 N 个文本句,并用向量空间模型表示;

(2) 使用本体中的一个或多个概念及其属性来对向量空间模型中的词进行权重的计算,得到句子的语义向量。可以得到 N 个语义向量: $s_i = (g_1, g_2, \cdots, g_w)$,其中, $i = 1, 2, \cdots, N$。w 为本体中的概念数;

(3) 根据公式(3)计算句子之间的相似度,结果用语义相似度矩阵表示;

(4) 确定聚类的阈值 T;

(5) 从语义相似度矩阵中选取相似度值最大的簇对 $\theta = \text{argmax} \sum Sim(s_i, s_j)$;

(6) 如果 $\theta < T$,表示语句类的相似度已经足够小,每个语

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

句类可以表示一个中心句。该算法结束;

(7) 将 s_i 和 s_j 合并成为一个新簇 $s_k = s_i \cup s_j$,生成文本的一个新的聚类:

$$C = (s_1, s_2, \cdots, s_i, \cdots, s_{N-1})$$

(8) 转到(5),迭代计算。

在计算过程中,可以通过调节 T 的取值大小来控制得到的聚类数量。T 的取值越小得到的聚类数量越少,当 $T=0$ 时,文章中的全部语句聚为一个类。T 取值越大,聚类数量越多,当 $T=1$ 时,文章中的每一个句子聚为一个类。

1.5 抽取代表句生成主题文本

上节中生成的聚类结果是构成主题的基础。聚类结果的每一类中包含了相似度较大的句子,这些句子的叙述的内容较为接近,它们叙述的是文章的同一个子主题,可以认为是重复叙述的内容。因此,在每类句子中选出重要度最大的一个句子,作为本类的代表句。要选出重要度最大的一个句子,就要计算聚类内每一类中所有句子两两之间的相似度,取与其它句子相似度之和为最大值的句子本类重要度最大的句子,既对于类:

$$c_i = (s_{i1}, s_{i2}, \cdots, s_{ij}, \cdots, s_{in})$$

其中 n 为类 c_i 包含的句子个数,代表句子为:

$$s^* = \underset{s_{ij}}{\operatorname{argmax}} \sum_{k \neq j} Sim(s_{ij}, s_{ik})$$

之后把这些句子按先后顺序排列起来,就生成了文章的主题。

1.6 基于本体的文本主题提取算法

按上述基本原理,提出基于本体的文本主题提取算法如下:

第一步:首先输入文本 T 和本体库;

第二步:对文本 T 进行切词和分句处理;

第三步:使用本体进行语义标注,得到句子的语义向量集合:$s_1, s_2, \cdots, s_N$,根据式(1)-式(3)计算语义相似度矩阵;

第四步:根据相似度矩阵对文本句进行聚类,聚类后得到一组语义上相对独立的类:

$$C_1, C_2, \cdots, C_i, \cdots, C_k$$

每一类包含语义上相近的句子;

第五步:从每个类中抽取重要度最大的一个句子作为该类的代表句,将代表句按一定的顺序连接起来,就得到文章的主题 $w_1, w_2, \cdots, w_i, \cdots w_k$,其中 w_i 是类 c_i 的代表句。

2 实验结果

实验采用国际语委语料库已经做了主题的 300 篇文章(包括新闻、军事、体育、教育、文学、经济六类)进行主题抽取。结果见表 1 所示。

表 1 测试语料集及测试结果

类别	篇数	速度	压缩比	评分结果
新闻	50	26s	17.3%	85
军事	50	25s	21.8%	90
教育	50	30s	24.2%	89
经济	50	28s	20.1%	83
体育	50	27s	23.5%	90
文学	50	29s	21.2%	84

实验结果表明,所产生的文章主题对原文本的信息压缩比较高,同时,所选取的代表句有较好的代表性,能够较好地描述文章的各个子主题。表明该系统生成的文本主题具有较好的可接受度。

3 结 语

文本主题的提取在信息检索、文本挖掘、Web 挖掘、图书情报管理、智能信息服务等领域将有广泛的应用前景。本文介绍了一种基于本体的文本主题提取方法的设计思想、整体结构和主要技术。在本文提出的系统中,将本体融入统计算法,该方法的特点是:先对文本进行切词和分句处理,再使用本体进行语义标注,得到句子的语义向量集合,然后进行语义相似度计算和语义聚类,最后从每类中抽取重要度最大的句子作为代表句,并以一定顺序连接代表句,组成文章主题。该方法原理简单,易于实现,具有较好的可扩展性,抽取的结果文本有较好的可接受度。实验结果表明,该方法是有效的,具有一定的实用价值。

参 考 文 献

[1] Jones K, Brigitte E N. Introduction: Automatic Summarizing [J]. Information Processing & Management, 2005, 31(5): 625-663.

[2] Hahn U, Mani I. The Challenges of Automatic Summarization [J]. IEEE: Computer, 2000.

[3] Jong G. Fast Skimming of News Stores: The FRUMP System [R]. New Haven: Yale University, 2002.

[4] Hahn U. The TOPIC Project: text-oriented procedures for information management and condensation of expository text [R]. University of Constance, W. Germany, 2005.

[5] Rau L F. Conceptual information extraction and retrieval from natural language input. User-Oriented Content-based Text and Image Handling [C]//M. I. T.: Proceedings of RIAO 88 Conference, 2008.

[6] Mc Keown K, Radev D. Generating summaries of multiple news articles [C]//Washington: Proceedings of ACM SIGIR'95. Seattle, 2005.

[7] Salton G, McGill M J. Introduction to Modern Information Retrieval [M]. New York: McGraw-Hill, 2003.

[8] Wang Hei-Chia, Liu Chia-Tzung, Huang Tian-Hsiang. A concept-level ontology construction method for automatic summarization [J]. Journal of Internet Technology, 2007, 8(4): 381-386.

[9] Lee Chang-Shing, Jian Zhi-Wei, Huang Lin-Kai. A fuzzy ontology and its application to news summarization [J]. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 2005, 35(5): 859-880.

[10] 李鹏,陶兰,王弼佐. 一种改进的本体语义相似度计算及其应用[J]. 计算机工程与设计, 2007, 28(1): 227-229.

(上接第 50 页)

[3] Zhang Deng-yin, Wang Qian-qian, Li Jun, et al. Novel multi-rate multicast congestion control algorithm [J]. Journal on Communications, 2007, 28(1): 135-140.

[4] Hu Ningning, Steenkiste P. Evaluation and Characterization of Available Bandwidth Probing Techniques [J]. IEEE Journal on Selected Areas in Communications, 2003, 21(6): 879-894.

[5] Claffy K, Miller G, Thompson K. The Nature of the Beast: Recent Traffic Measurements from an Internet Backbone [C]//Proc. of ISOC INET Conference. NY, USA, 1998.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net