基于 GVSM 的文本相似度算法研究

郑小波,郑 诚,尹莉莉

(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:提出了一种基于 WordNet 和 GVSM 的文本相似度算法,通过语义的路径长度和路径深度计算两个词的语义相似度,结合改进的 GVSM 模型计算文本相似度,并对基于 TFIDF-VSM 模型和本文方法进行了比较。实验结果表明,该算法取得了更好的准确率和效率。

关键词:文本相似度;语义相似度;词网;广义向量空间模型

中图分类号: TP391 **文献标识码**: A **文章编号**: 1674-7720(2011)03-0009-03

Research on similarity algorithm of text based on GVSM

Zheng Xiaobo, Zheng Cheng, Yin Lili

(Key Lab. of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University, Hefei 230039, China)

Abstract: This paper presents a text similarity algorithm based on WordNet and GVSM, computing the similarity of two words by semantics of path length and depth, combined with the improved GVSM model. Then compare the TFIDF-VSM-based model with this method. The experimental results show that this algorithm can achieve a better precision and efficiency.

Key words: text similarity; semantic relatedness; WordNet; GVSM

文本相似度计算在文本信息处理相关领域有着广泛的应用。目前,文本相似度的研究主要有三种方式:(1)篇章与篇章之间的相似度计算[1];(2)短语与篇章之间的相似度计算;(3)短语与篇章中段落的相似度计算。文本相似度计算方法主要有隐性语义索引模型、向量空间模型、广义向量空间模型、基于属性论的方法、基于海明距离的计算方法、基于数字正文的重构方法等。基于语义的相似度计算方法相关的研究主要有:使用 WordNet 进行相似度计算的方法;使用同义词词林进行相似度计算的方法[2];使用知网《HowNet》知识结构进行相似度计算的方法[3]。广义向量空间模型(GVSM) 是 20 世纪 80 年代由 Wong 提出[4],在词语消歧研究[1]、文本检索研究[5]等方面得到了很好的应用。

本文使用 WordNet 进行相似度计算的方法,采用广义向量空间模型,并对广义向量空间模型进行了扩展,得到了新的广义向量空间模型。通过 WordNet 计算两个词的语义相似度,把语义相似度应用到 GVSM 模型中来计算文本相似度。实验结果表明,该算法取得了较好的准确率和效率。

1 背景知识介绍

1.1 向量空间模型

向量空间模型(VSM)是 20 世纪 70 年代末由 Salton 等[6]提出的一种代数模型。在近 30 年内,向量空间模型(VSM)被广泛应用到信息检索、文本分类、文本聚类等领域,并取得了很好的效果。其基本思想是:假设词与词之间是不相关的,以向量表示文本,每个维度对应于一个单独的词,则$(w_1, w_2, w_3, \cdots, w_n)$文档 d_k 可以看成相互独立的词条$(t_1, t_2, t_3, \cdots, t_n)$,为了表示词条的重要程度,给每个词条赋予相应的权值 w_i,其中文档 d_k 可用向量$(w_1, w_2, w_3, \cdots, w_n)$表示。向量空间模型中的文档相似度计算方法为:

$$\mathrm{sim}(d_k, d_p)=\frac{\sum_{i=1}^{n} w_{ki}\times w_{pi}}{\sqrt{\sum_{i=1}^{n} w_{ki}^2}\times\sqrt{\sum_{i=1}^{n} w_{pi}^2}} \quad (1)$$

其中 w_{ki}、w_{pi} 分别是词 t_i 在 d_k 和 d_p 的权值,n 是向量的维度。向量空间模型的前提是假设词与词之间是不相关的,但这种假设不现实,因为词与词之间往往存在语义相关。

1.2 广义向量空间模型

广义向量空间模型 GVSM 扩展的 VSM 模型，GVSM 引入了词与词之间的相关度，并提出了一个新的向量空间，每个向量 t_i 被表示成 $2n$ 维向量 m_r，其中 $r=1,2,\cdots,2n$。文档相似度计算方法为：

$$\mathrm{sim}(d_k,d_p)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ki}\times w_{pi}\times R(t_i,t_j)}{\sqrt{\sum_{i=1}^{n}w_{ki}^2}\times\sqrt{\sum_{i=1}^{n}w_{pi}^2}} \tag{2}$$

其中 w_{ki}、w_{pi} 分别是词 t_i 在 d_k 和 d_p 的权值，$R(t_i,t_j)$ 是词 t_i 和 t_j 的相关度。

1.3 WordNet 介绍

WordNet 由普林斯顿大学认知科学实验室在 1985 年建立，是一部在线词典数据库系统，采用了与传统词典不同的方式，即按照词义而不是词形来组织词汇信息。WordNet 将英语的名词、动词、形容词、副词组织为 Synsets，每一个 Synset 表示一个基本的词汇概念，并在这些概念之间建立了包括同义关系(synonymy)、反义关系(antonymy)、上下位关系(hypernymy & hyponymy)、部分关系(meronymy)等多种语义关系。不同的边代表不同的语义关系。

2 文档相似度计算

2.1 语义相似度计算

本文模型中使用 WordNet 衡量两个词的语义关系。分别考虑了路径长度 *SPC*(Semantic Path Compactness)和路径深度 *SPE*(Semantic Path Elaboration)，给定两个词的语义相关度 *SR*(Semantic Relatedness)由 *SPC* 和 *SPE* 合并得出。下面给出相关定义。

定义 1：给定一个词库 O、一组词义 $S=(s_1,s_2)$和一条 s_1 到 s_2 路径 l，并对于每条边进行加权处理，其中权值 $e\in(0,1)$，则 *SPC* 定义为：

$$SPC(S,O)=\prod_{i=1}^{l}e_i \tag{3}$$

其中 $e_1,e_2,e_3,\cdots,e_l$ 分别是每条边的权值；当 $s_1=s_2$ 时，$SPC(S,O)=1$；如果 s_1 与 s_2 之间没有路径，则 $SPC(S,O)=0$。

定义 2：给定一个词库 O、一组词义 $S=(s_1,s_2)$和一条 s_1 到 s_2 路径 l，其中 $s_1,s_2\in O$ 且 $s_1\neq s_2$，则 *SPE* 可定义为：

$$SPE(S,O)=\prod_{i=1}^{l}\frac{2d_i\cdot d_{i+1}}{d_i+d_{i+1}}\cdot\frac{1}{d_{\max}} \tag{4}$$

其中 d_i 是 s_i 的深度，$d_{\max}$ 是 O 的最大深度。当 $s_1=s_2$ 且 $d=d_1=d_2$ 时，$SPE(S,O)=\dfrac{d}{d_{\max}}$；如果 s_1 与 s_2 之间没有路径，则 $SPE(S,O)=0$。

定义 3：给定一个词库 O、一组词 $T=(t_1,t_2)$和它们所有的词义 $S=(s_{1i},s_{2j})$，其中 s_{1i} 和 s_{2j} 分别是 t_1 和 t_2 的词义，则 $SR(T,S,O)$可定义为：

$$SR(T,S,O)=\mathrm{MAX}\{SPC(S,O)\times SPE(S,O)\} \tag{5}$$

其中 $T=(t_i,t_j)$，$i=j=1,2,\cdots,n$。当 $t=t_i=t_j$ 时，$SR(T,S,O)=1$；当 $t_i\in O$ 且 $t_j\notin O$ 或 $t_i\notin O$ 且 $t_j\in O$ 时，$SR(T,S,O)=0$。

2.2 语义网络构建

为了计算两个词的语义关联度，需要构建语义网络，采用了文献[7]的方法。相比较其他方法，它嵌入所有可用的 WordNet 的语义信息并提供了丰富的语义表达。根据所采用语义网络建设模式，每种类型的边将被赋予各自的权值，权重越高说明它们的语义关联度越高(如上位/下位边的权值定义为 0.57)。词与词义的关系在语义网中如图 1 所示。

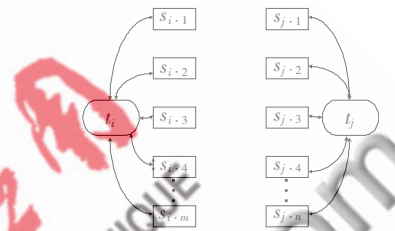

图 1 词与词义关系

其中 $s_{i\cdot m}$、$s_{j\cdot n}$ 分别是词 t_i 和 t_j 的词义，m 是词 t_i 的词义数，n 是词 t_j 的词义数。

遍历 t_i 和 t_j 所有的词义，将会出现以下几种情况：

(1)如果 $s_{i\cdot m}$ 和 $s_{j\cdot n}$ 之间没有路径，如图 2(a)所示，则 $SR((t_i,t_j),(s_{i\cdot m},s_{j\cdot n}),O)=0$。

(2)如果 $s_{i\cdot m}$ 和 $s_{j\cdot n}$ 之间只有一条路径，如图 2(b)所示，则 $s_{i\cdot m}$ 和 $s_{j\cdot n}$ 的语义关联度为 $SPC((s_{i\cdot m},s_{j\cdot n}),O)$。

(3)如果 $s_{i\cdot m}$ 和 $s_{j\cdot n}$ 之间有多条路径，如图 2(c)所示，则 $s_{i\cdot m}$ 与 $s_{j\cdot n}$ 的语义关联度为 $\max\{SPC((s_{i\cdot m},s_{j\cdot n}),O)\times SPE((s_{i\cdot m},s_{j\cdot n}),O)\}$。

(4)如果 t_i 和 t_j 有两个相同的词义，如图 2(d)所示，则 $SR((t_i,t_j),(s_{i\cdot m},s_{j\cdot n}),O)=\dfrac{d}{d_{\max}}$。

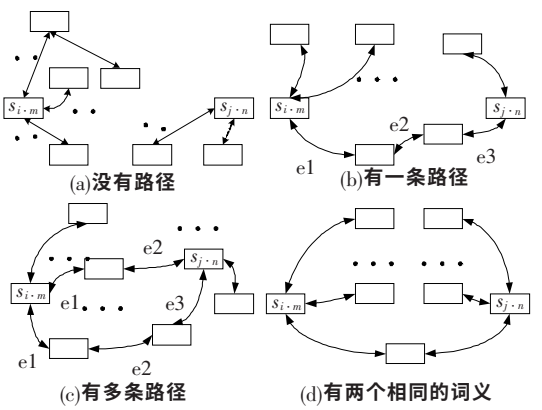

图 2 语义关系

2.3 文本相似度计算

式(2)中介绍了 GVSM 模型，现将式(5)应用到 GSVM 模型中，使得：

$$R(t_i,t_j)=SR((t_i,t_j),(s_{i\cdot m},s_{j\cdot n}),O) \tag{6}$$

这里定义一个新的文本向量，新向量中增加了 t_i 和 t_j 在文本 d_k 中的 *TF-IDF* 权值，如下定义：

$$d_k(t_i,t_j)=(tf-idf(t_i,d_k)+tf-idf(t_j,d_k))\cdot R(t_i,t_j) \quad (7)$$

由新的文本向量可以产生一个新的 GVSM 模型，则两个文本之间的相似度公式定义为：

$$\mathrm{sim}(d_k,d_p)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}d_k(t_i,t_j)\cdot d_p(t_i,t_j)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}d_k(t_i,t_j)^2}\times\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}d_p(t_i,t_j)^2}} \quad (8)$$

其中 n 为向量的维度，d_k 和 d_p 分别是两篇不同的文档。

3 实验

利用上述方法，本文实现了基于 WordNet 的语义相似度计算程序模块。为了对相似度计算结果更好地进行分析，本文评价的方案放在文本分类系统中，以观察不同计算方法对文本分类系统性能的影响。

3.1 实验评价标准

评价标准是在测试过程中所使用的一些用来评价分类器分类准确度的量化标准。本文采用常用的三种标准，它们在不同的方面来评价一个分类器。

准确率(precision)= (分类正确的文本数)/(实际分类的文本数)

召回率(recall)= (分类正确的文本数)/(应有分类正确的文本数)

$$F1=\frac{2pr}{p+r}\quad (p\text{ 为准确率},r\text{ 为召回率})$$

3.2 实验结果与分析

本文实验是在 Windows XP 操作系统、Eclipse 开发环境下，通过 Java 语言实现。实验是在 1 GB 内存、P4 3.0 GHz CPU 的 PC 机下进行的。实验数据集采用的是 20-Newsgroups 文本数据集。20-Newsgrops 是在 UseNet 上下载的 20 个类的新闻组讨论英文文章。数据集共有 20 个类，每个类大约 1 000 篇。20-Newsgroups 是一个比较常用的文本数据集。出于效率考虑，本实验选取其中的 5 个类别，针对不同数量的训练文本进行了实验，实验分别选取了 200、400、600、1 000、2 000 篇文本平均分配到编号为 A、B、C、D、E 的 5 个集合。分别对基于 TFIDF-VSM[3]模型和本文提出的基于 WordNet 的 GVSM 模型进行了比较实验。本文采用 KNN[8]分类器进行评价，测试结果记录了上述 5 种情况分类器的准确率、召回率、*F*1 值。

实验结果表明，采用基于 WordNet 的 GVSM 模型比基于 TFIDF-VSM 模型具有更高的准确率、召回率、*F*1 值。分析发现当文本数越多时，文本分类的准确率、召回率、*F*1 值越高。

本文提出了一个新的文本相似度计算方法，将其成功地应用在文本分类当中，实验证明得到了很好的效果。首先基于 WordNet 构建了语义网，分别考虑路径长度 *SPC* 和路径深度 *SPE* 来计算两个词的语义关联度；然后将其应用在 GVSM 模型中计算文本相似度；最后应用在文本分类中，得到了较高的分类准确率和召回率。下一步准备将其应用到信息检索中，以提高信息检索的准确率与效率。

参考文献

[1] WILLETT P. Recent trends in hierarchical document clustering: a critical review. Inf Process and Manage, 1988:577-597.

[2] 夏天. 汉语词语语义相似度计算研究 [J]. 计算机工程, 2007, 33(6): 191-194.

[3] 李峰, 李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报, 2007, 21(3): 99-105.

[4] WONG, S. K. M. Wojciech Ziarko, Patrick C. N. Wong. Generalized vector spaces model in information retrieval. SIGIR ACM, 1985.

[5] TSATSARONIS G, PANAGIOTOPOULOU V. A generalized vector space model for text retrieval based on semantic relatedness. Proceedings of the EACL 2009 Student Research Workshop, 2009:70-78.

[6] SALTON, MCGILL M J. Introduction to modern information retrieval. McGraw-Hill, 1983.

[7] VAZIRGIANNIS T M. Word sense disambiguation with spreadingactivation networks generated from thesauri [C]. In Proc. of the 20th IJCAI, 2007:1725-1730.

[8] HALL P, PARK B U, SAMWORTH R J. Choice of neighbor order in nearest-neighbor classification. Annals of Statistics: 2008:2135-2152.

[9] Qinglin Guo. The similarity computing of documents based on VSM. IEEE International Computer Software and Applications Conference. 2008:585-586.

(收稿日期：2010-09-12)

作者简介：

郑小波，男，1983 年生，硕士研究生，主要研究方向：信息检索与文本数据挖掘。

郑诚，男，1964 年生，副教授，硕士生导师，主要研究方向：数据挖掘、机器学习研究。

尹莉莉，女，1985 年生，硕士研究生，主要研究方向：数据挖掘。