一种基于本体的句子相似度计算方法

刘宏哲

(北京联合大学北京市信息服务工程重点实验室 北京 100101)

摘 要 提出了一种基于树结构本体的句子相似度计算方法。利用本体概念与句子中关键词之间建立的语义索引,构建句子与本体间的直接和间接语义联系,据此提取描述句子的语义向量,从而计算句子间的语义相似度。应用微软研究院的意译语料库(MSRP)对本方法进行了验证,结果表明:与相关的计算方法相比,本方法在不完备附加信息应用前提下获得了较好的准确率和召回率。

关键词 句子相似度计算,本体,WordNet

中图法分类号 TP391 **文献标识码** A

Ontology Based Sentence Similarity Measurement

LIU Hong-zhe

(Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China)

Abstract This paper proposed sentence similarity computing based on ontology. Using the relations between the ontology concepts and key words in the sentences to establish semantic index to extract the direct and indirect semantic relation, ontology based semantic vector was represented to calculate the semantic similarity between sentences, thus the sentence similarity computing method was proposed. This method is applied in the Microsoft Research Institute of paraphrase corpus (MSRP). Experiments show that compared with the related similarity computing methods, this method obtains good accuracy and recall rate in the incomplete additional information background.

Keywords Sentence similarity computing, Ontology, WordNet

1 简介

通常情况下,如果句子对有相同意思或者主旨一致,那么就认为该句子对是相似的。许多自然语言处理应用要求简短文字段落或句子间的相似度能够快速、准确地计算出来。一种能自动计算语义相似度的方法在自动问答[1,2]、信息过滤[3]、文献摘要[4]、机器翻译[5]等方面十分有价值。除了基于表面意义的字重叠(Word Overlap)句子相似度方法外,大多数现有的方法不仅依赖于像 WordNet 这样的本体知识库,而且依赖大的语料库作为额外的知识资源。然而,在许多应用中,尤其是在基于领域的应用中,大规模语料库不能随时获取到。在这种情况下,这些简短段落或句子之间的相似度只能从有限的表述中提取。本文研究如何仅通过本体结构所表达出来的概念间的语义关系来计算句子的相似度。

2 相关研究

2.1 相关方法分类

根据相关文献[6,7],目前可用于确定句子相似度的方法主要有4类,分别是基于字重叠(Word Overlap)的方法、基于语料库统计(TF-IDF)的方法、基于语言学(Linguistic)的方法和混合方法,如表1所列。

表1 相关方法分类

方法分类	类描述	主要方法	方法描述
字重叠方法 (Word Overlap Measures)	一组通过两个句子所共有的一些词汇量来计算句子的相似度度量方法	Jaccard 相似系数法[8]	两句子中词语交集与两句子中词语并集的比值
		简单词汇重叠法[9,10]	利用句子的长度使两个句子中都出现的词语的比例归一化
		IDF 重叠法[9]	用逆向文档频率(IDF)作为两个句子中均出现的词语的权重
		Zipfian 重叠法[11,12]	基于短语的长度和它们的使用频率呈 Zipfian 分布的特点来设计基于短语的句子相似度计算方法
基于语料库的方法 (Corpus based Measures)	把句子对中出现的词语集合用来作为特征集,将基于语料库的向量的余弦夹角值作为相似值	LSA[13]	通过分析一个大型的自然语言语料库来统计关键词的 TF-IDF 值形成句子语义向量,用向量的余弦夹角来计算句子语义相似度
		HAL[14]	统计词汇之间的共现性得到高维向量空间来计算句子或短文档相似度
		Allan 的 TF-IDF 法[10]	一种计算在句子对中共同出现的词语和的方法,词语用 TF * IDF 作为权重

到稿日期:2012-04-11 返修日期:2012-08-01 本文受国家自然科学基金项目(60972145),北京市教委科技面上项目(KM201111417002),北京市属高等学校人才强教计划项目(PHR201108419)资助。

刘宏哲(1971—),女,博士,副教授,主要研究方向为语义计算、人工智能、数字博物馆,E-mail:xxtliuhongzhe@buu.edu.cn。

续表

方法分类	类描述	主要方法	方法描述
基于语言学方法 (Linguistic Measures)	利用词汇间的语义关系及其语法成分来确定句子的相似度	Li 的方法[16]	一种语义向量相似度和词序相似度的线性组合方法
		Mihalcea 的方法[17]	基于词语语义相似度度量句子间的相似度，考虑单词具有不同的区分能力来形成句子向量的相似度计算方法
		Malik 的方法[18]	组成句子对的词之间的相似度的总和的最大值被句子长度归一化所得值作为句子相似度值
改进或混合方法 (Improved or hybrid Measures)	基于以上方法的混合方法，或者利用以上方法的思想具体应用	混合方法[7,15,19-27]	基于以上方法在中文句子相似度计算上的应用，其它方法属于混合方法

2.2 相关方法分析

句子相似度方法可以分为：字重叠方法(Word Overlap Measures)，主要通过两个句子所共有的一些词汇量来计算句子的相似度；基于语料库的方法(Corpus based Measures)，把句子对中出现的词语集合作为特征集，将基于语料库统计的向量的余弦夹角值作为相似值；基于语言学方法(Linguistic Measures)，主要利用词汇间的语义关系及其语法成分来确定句子的相似度；最后就是改进方法或者混合方法(Improved or hybrid methods)，主要是基于以上方法的混合方法，或者利用以上方法的思想具体应用。

通过对上面 3 种方法的比较发现，字重叠方法相对来说比较简单，然而在短文本中，由于自然语言所固有的灵活性，人们可以用不同结构或单词内容的句子来表达同样的意思，因此通过句子的表面相似度来判断句子的相似度并不可靠[6]。基于 TF-IDF 统计的语料库方法需要一个大文本语料库作统计计算，而对于某些领域数据，这样一个文本语料库很难在某一个具体领域中找到。基于语言学方法在准确度上比其它方法要强，但是该方法的运算基础是首先算出句子对中词对词的相似度，所以性能相对来说比较低。例如对于两个含有 50 个单词的句子，这类方法需要访问 WordNet 结构 50 * 50 次来计算词义相似度，所以相应地要花费至少 2500 次的词语相似度计算来完成一个句子对的相似度计算。

3 基于本体的句子相似度计算方法

本文以本体为知识框架，概念之间语义关系通过本体结构来体现，通过发现本体术语和句子内容之间的联系生成句子与本体之间的索引，并把这种联系运用到计算句子相似度的方法中，最终形成一个高效的句子相似度计算方法。

3.1 相关定义

首先，本文方法的前提是一个基于树的本体结构，该结构的本体是本文所提出的方法的基础(见定义 1)。

定义 1(概念层次树) HCT 可表示为 $T(N,E)$，N 是树中的概念节点集，而 E 是 H 中父/子对之间的连接集，子概念节点覆盖的语义范围是其父概念节点覆盖范围的一部分，且它们之间相互独立。

在 HCT 的基础上，相似度的计算取决于余弦相似度，要求其组成部分间相互垂直，所以概念节点覆盖的语义范围应该是独立的。即同级概念节点包含的语义通常是非重迭的；只有通过其父概念节点来捕获两个兄弟概念节点之间的关系。

在 HCT 结构中，任何给定节点的祖先概念节点都包含了该概念节点的属性，而它的派生概念节点也继承了它的属性。所以可以说祖先和派生概念节点都与这个概念节点相关。定义 2 是本文给出的相关概念节点的定义。

定义 2(相关概念节点) 在 HCT 中给定的概念节点的相关概念节点是其祖先和派生概念节点的节点集合。

定义 2 是本文创新工作的基础，是将基于表面意义相似度计算转化为基于语义概念相似度计算的基础。

在进行句子相似度计算之前，首先对句子进行预处理。一个中文或者英文句子可能由许多词组成，这些词中包括冠词、介词、连词、助词等等。这些词对于句子的语义不甚重要，所以本文在进行相似度计算的时候将这些词丢弃。剩下的则是句子的关键词，本文将句子中的这些关键词映射到领域本体中。在定义 3 和定义 4 中定义了句子和本体间的直接连接概念节点和间接连接概念节点。

定义 3(句子直接连接概念节点) 如果句子中的某个关键词等同于 HCT 中的概念节点，那么该概念节点就叫做这个关键词的直接连接概念节点。所有句子的关键词在本体中的直接连接概念节点的集合就叫做该句子的直接连接概念节点。

定义 4(句子间接连接概念节点) 句子的间接连接概念节点是指每一个句子的直接连接概念节点在 HCT 结构中的相关概念节点的并集。

句子间接连接概念节点的发现是本创新工作的关键，它是将基于表面意义的句子相似度计算转化为基于语义的相似度计算的基础。定义 5 定义了如何基于 HCT 结构中直接连接概念节点和间接连接概念节点而形成一个句子向量。

定义 5(句子向量) 给出一个有 n 个概念节点的 HCT，其中句子的概念向量定义为 $\vec{S}=(v_1,v_2,\cdots,v_n)$，$v_i$ 是维度值，i ($i=1,2,\cdots,n$)是维数，维度值定义如式(1)所示：

$$v_i=\begin{cases}1, & C_i \text{ 是句子的直接连接节点}\\ w, & C_i \text{ 是句子的间接连接节点}(0<w<1)\\ 0, & \text{其它}\end{cases} \quad (1)$$

对于相似度计算而言，由于本体中句子的间接连接概念节点的重要性比句子直接连接概念节点低，因此本文赋予其较小的权重 $w(0<w<1)$。

3.2 方法的工作步骤

图 1 中描述了基于 HCT 结构的领域本体，假设两个句子包含的关键词集合为 T_1 和 T_2。

$T_1=\{S_1 \text{ 中的关键词}\}=\{word_1,word_2\}$

$T_2=\{S_2 \text{ 中的关键词}\}=\{word_a,word_b\}$

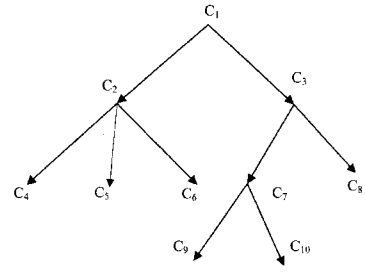


图 1 HCT 结构的领域本体

本文通过 4 步来计算句子相似度：

第 1 步　将句子中的关键字与本体中概念相映射来查找句子直接连接概念节点。

如图 2 可知，句子 1 中的 $word_1$、$word_2$ 映射到本体中的概念节点 C_3、C_4，根据定义 1 知，C_3、C_4 是句子 1 的直接连接概念节点；句子 2 中的 $word_a$、$word_b$ 映射到本体中的概念节点 C_2、C_9，则 C_2、C_9 是句子 2 的直接连接概念节点。T_1 和 T_2 表示如下：

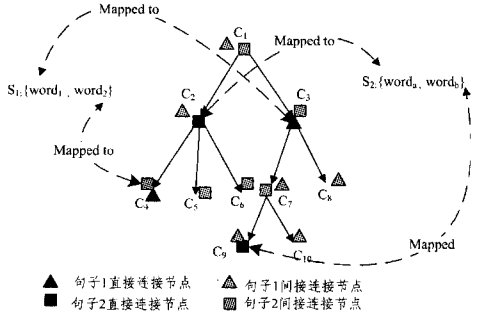


图 2　直接连接概念节点映射

$$S_1=\{\text{句子 1 中的关键字}\}=\{word_1,word_2\}\xrightarrow{\text{映射}}\{C_3,C_4\}$$

$$S_2=\{\text{句子 2 中的关键字}\}=\{word_a,word_b\}\xrightarrow{\text{映射}}\{C_2,C_9\}$$

第 2 步　根据直接连接概念节点在本体中的相关概念节点，扩充句子的直接连接概念节点至直接连接概念节点和间接连接概念节点的并集。

根据定义 2，C_3 的相关概念节点（$Rel(C_3)$）有 C_1、C_7、C_8、C_9、C_{10}；C_4 的相关概念节点（$Rel(C_4)$）有 C_1、C_2。根据定义 4，句子 S_1 与本体的间接连接概念节点为 $Rel(C_3)\cup C_4(Rel(C_4))$。

同样根据定义 2，C_2 的相关概念节点（$Rel(C_2)$）为 C_1、C_4、C_5、C_6、C_9 的相关概念节点（$Rel(C_9)$）为 C_1、C_3、C_7，根据定义 4，句子 S_2 与本体间的间接连接概念节点为 $Rel(C_2)\cup C_4(Rel(C_9))$。将 S_1 和 S_2 与本体间的直接相关概念节点扩展到直接连接概念节点和间接连接概念节点的并集：

$$\begin{aligned}S_1&=\{\text{句子 1 中的单词}\}\\&=\{word_1,word_2\}\xrightarrow{\text{映射}}\{C_3,C_4\}\xrightarrow{\text{扩展}}\{C_3,C_4,C_1,C_2,C_7,C_8,C_9,C_{10}\}\\S_2&=\{\text{句子 2 中的单词}\}\\&=\{word_a,word_b\}\xrightarrow{\text{映射}}\{C_2,C_9\}\xrightarrow{\text{扩展}}\{C_2,C_9,C_1,C_3,C_4,C_5,C_6,C_7\}\end{aligned}$$

从上面的例子可以看到，如果只考虑从句子关键词到本体的直接映射，句子 S_1 和 S_2 之间并没有任何交集，但是当考虑到间接连接概念节点时，它们之间就存在交集了。这一步是将基于字面重合的表面语义相似度计算转化成基于真正语义的相似度计算方法的关键之处，也是本文创新工作的关键所在。

第 3 步　形成句子 S_1、S_2 基于领域本体的概念向量。

与直接连接概念节点所形成的语义关系相比，由间接连接概念节点所形成的语义关系则有些“浅”。所以本文给予间接连接概念节点相对较小的权重 $w(0<w<1)$。图 1 作为句子 S_1 和 S_2 的领域本体，根据定义 3，形成句子 S_1 和 S_2 基于直接连接概念节点和间接连接概念节点的句子概念向量。如果根据树的广度优先遍历序列列出所有概念节点在概念向量中的出现次序，则有如下概念向量：

$\vec{S_1}=(w,w,1,1,0,0,w,w,w,w)$

$\vec{S_2}=(w,1,w,w,w,w,w,0,1,0)$

第 4 步　用向量余弦相似度公式计算句子相似度。

句子 S_1 和 S_2 之间的语义相似度 $\text{sim}(S_1,S_2)$ 计算如式(2)所示。

$$\text{sim}(S_1,S_2)=\frac{\vec{S_1}\cdot\vec{S22}}{\|\vec{S_1}\|\ \|\vec{S_2}\|}\tag{2}$$

3.3　实验评估

本文分别基于领域本体（中国青铜器）、通用数据集“微软研究院释义语料库（MSRP）”数据集合（句子测试集）[28] 和 WordNet（本体）[29] 进行句子相似度计算实验，并将本文方法与其它相关方法进行了比较分析。

3.3.1　基于领域本体计算句子相似度

实验步骤：

图 3 是中国青铜古董分类，这些边对应为“is a”关系。相关方法中唯一能够应用的方法就是基于词重叠的方法。

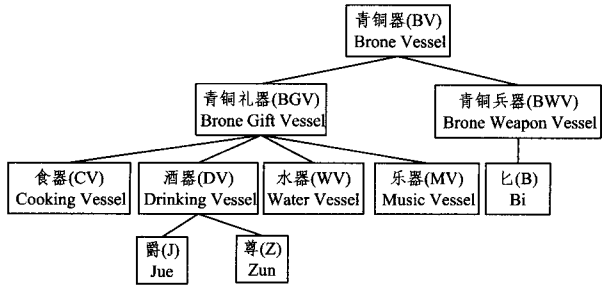


图 3　中国青铜古董分类

用于计算句子相似度的两个句子列举如下：

句子 1：CV is used as cooking ware in old times of china.

句子 2：Ancient Chinese use the BGV to cook, drink, and play music.

句子 S_1 和句子 S_2 所描述内容相似，所以根据人们的判断该句子对应该是相似的。接下来，将使用本文方法计算该句子对的相似度。

根据第 1 步和第 2 步，有

$S_1=\{S_1\text{ 中的关键词}\}=\{\text{CV, cooking, ware, old, times, china}\}\xrightarrow{\text{映射}}\{\text{CV}\}\xrightarrow{\text{扩展}}\{\text{CV, BGV, BV}\}$

$S_2=\{S_2\text{ 中的关键词}\}=\{\text{ancient, Chinese, use, BGV, cook, drink, play, music}\}\xrightarrow{\text{映射}}\{\text{BGV}\}\xrightarrow{\text{扩展}}\{\text{BGV, BV, CV, DV, WV, MV, J, Z}\}$

根据第 3 步，图 3 中基于领域本体的两个句子的概念向量表示如下：

$\vec{S_i}=(v_1,v_2,\cdots,v_n)$

- 根据图 3 中领域本体概念节点的数目，$n=10$。
- $v_1,v_2,\cdots,v_n$ 的连续顺序是根据图 3 中的结构广度优先遍历序列所给出的。

所以句子对的概念向量表示如下：

$\vec{S_1}=(w,w,0,1,0,0,0,0,0,0)$

$\vec{S_2}=(w,1,0,w,w,w,w,0,w,w)$

句子对之间的相似度用两个对应向量的余弦夹角来计算(假设权重 $w=0.5$):

$$\mathrm{sim}(S_1,S_2)=\frac{\vec{S_1}\cdot\vec{S_2}}{\|\vec{S_1}\|\ \|\vec{S_2}\|}=0.62 \tag{3}$$

实验结果分析:

表 2 是运用各个方法的实验结果比较。

表 2 结果比较

对比	结果	分析
Jaccard 相似度系数[8]	0	与人们的判断不符合
简单字重叠部分[9]	0	与人们的判断不符合
IDF overlap[9]	0	与人们的判断不符合
Zipfian overlap[11]	0	与人们的判断不符合
本文方法	0.61	与人们的判断符合
语料库统计方法类	没有大型文本语料库,无法使用该方法	
语言学方法类	没有字典注释、文本语料库等,无法使用该方法	

大部分现有的方法无法应用于图 3 中的古董分类本体,唯一适用的方法是词重叠方法。而在能够应用的词重叠方法中,Jaccard 相似度系数[8]、简单字重叠部分法[9]、IDF 重叠法[9]和 Zipfian 重叠法[11]都是基于句子对间是否具有一定数量的相同的词(或者短语)来判断两个句子语义是否等效。本文测试的基于中国青铜器本体的句子对之间不存在任何相同的词或者短语,所以计算出的结果都是"0"。基于本文提出的方法能算出的相似度值为 0.61,最好地反映了客观情况。而且这一实验表明,在构建一个新的应用时,本文方法具有很好的适应性。

3.3.2 标准数据集测试本文方法

为了对提出的方法进行全面评估,本文在文献[6]中引入了 3 个分类中的 14 种句子相似度方法来与本文方法进行比较。

数据集

微软研究院释义语料库(MSRP)数据集是一个人工衡量句子相似度计算的资料组。这个资料组包含取自互联网新闻文章的 4076 个训练句子对和 1,725 个测试句子对[28]。每个句子对都由两位评估人员来判断其是否语义等效。总体而言,3900 对句子(占所有句子的 67%)被判断是语义等效的。为了与相关方法进行试验比较,本文同样将 1725 个测试句子对作为测试目标。语义等效句子对可能包含完全相同的信息,或者包含细微差别的信息。描述同一事件但却是另一句子的超集的句子,被认为是不等同的句子对。MSRP 的简单统计情况如表 3 所列。

表 3 数据集情况介绍

数量	MSRP 数据集(对)	训练句子集(对)	测试句子集(对)
总数	5801	4076	1725
语义等效	3900	2753	1147
语义不等效	1901	1323	578

本体

该实验使用 WordNet 作为通用本体,WordNet 中的名词和动词构成的树作为句子映射的基础。名词中的 hyponymy/hypernymy 和动词中的 Hypernymy/ Troponymy 关系都属于上位/下位关系,是占绝大多数的基本关系。如图 4 所示,虚线部分是人为加入的边,目的是将名词和动词组成一棵 HCT 树统一处理。严格地说,基于 WordNet 中名词和动词的 HCT 不算一个十树结构,因为有些概念节点不止一个父概念节点。本文随机地选择其中的一个父概念节点来生成概念向量(实验表明,选择不同的父概念节点对实验结果影响很小)。

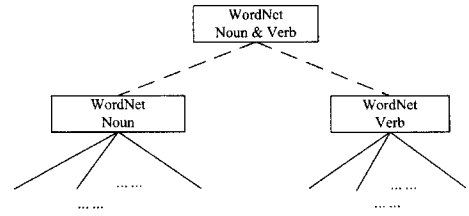


图 4 WordNet 名词和动词树

评估标准

本工作考虑了 4 种不同的评价措施:准确度(Accuracy)、精度(Precision)、召回率(Recall)以及 F-度量值(F-measure)来衡量方法的性能,这 4 个评价措施的定义如下:

$Accuracy=(TP+TN)/(TP+TN+FP+FN)$

$Precision=TP/(TP+FP)$

$Recall=TP/(TP+FN)$

$F\text{-}measure=(1+\beta)PR/(\beta P+R)=2PR/(P+R)$(当 $\beta=1$ 时,*Precision* 和 *Recall* 有同样的权重)

TP:预测为相似并且实际上也是相似的句子的数量;

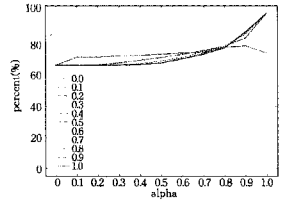
TN:预测为不相似并且实际上不相似的句子的数量;

FP:预测为相似但实际上不相似的句子的数量;

FN:预测为不相似但实际上相似的句子的数量。

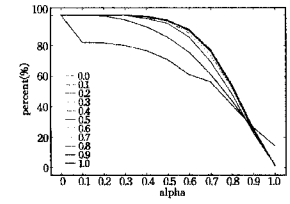
实验结果分析

图 5 和图 6 分别显示了在不同相似度阈值(*alpha*)和不同间接连接概念节点权重(w)下的精度(Precision)召回率(Recall)。实验得出,当间接连接概念节点权重 $w\geqslant 0.4$ 时,通过方法得到的精度、召回率有相近的曲线形状。



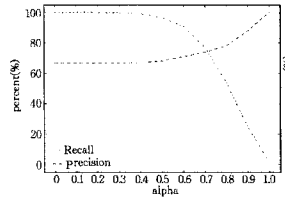
X 轴:相似度阈值(alpha)
Y 轴:精度值 (percent(%))

图 5 在不同相似度阈值(alpha)和不同间接连接概念节点权重(w)下的精度



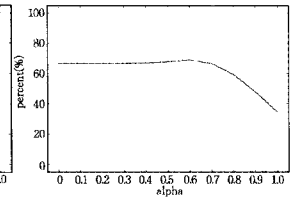
X 轴:相似度阈值(alpha)
Y 轴:精度值 (percent(%))

图 6 在不同相似度阈值(alpha)和不同间接连接概念节点权重(w)下的召回率



X 轴:相似度阈值(alpha)
Y 轴:精度值 (percent(%))

图 7 在间接连接概念节点权重(weight $w=0.6$)时不同相似度阈值(alpha)下的精度和召回率



X 轴:相似度阈值(alpha)
Y 轴:精度值 (percent(%))

图 8 在间接连接概念节点权重($w=0.6$)时不同相似度阈值(alpha)下的准确度

图 7 所示的是间接连接要领节点权重 $w=0.6$ 时,不同

相似度阈值下的精度和召回率，图 8 所示的是当间接连接概念节点权重 $w=0.6$ 时，不同相似度阈值下的准确度；可以看出当相似度阈值 $alpha=0.6$ 时的最高准确度值是 0.69。表 4 列出了相关各类方法的精度(Precision)、召回率(Recall)、准确度(Accuracy)和 F-度量值(F-measure)的平均值。

表 4 相关各类方法的 Precision，Recall，Accuracy 和 F-measure 的平均值

方法	精度 (Precision)	召回率 (Recall)	准确度 (Accuracy)	F-measure
基于字重叠的方法	0.78	0.62	0.62	0.66
基于语料库方法	0.74	0.75	0.63	0.69
基于语言学的方法	0.68	0.89	0.65	0.76
本文方法 (w=0.6，alpha=0.6)	0.71	0.91	0.69	0.80

其中，前面 3 种方法的精度(Precision)、召回率(Recall)、准确度(Accuracy)和 F-度量值(F-measure)的平均值是根据参考文献[6]中的方法来计算的。

本文根据准确度(Accuracy)、精度(Precision)、召回率(Recall)以及 F-度量值(F-measure)来分析计算结果。召回率是正确预测出的相似句子占所有相似句子的比例。精度是正确预测出的相似句子占所有预测为相似的句子的比例。大多数方法能得到较好的精度或召回率，但很少能使两者同时获得较好的数值。准确度是所有正确预测出的句子占所有句子的比例，准确度值越高，代表实验结果越准确。根据表 4 可知，基于字重叠方法的平均准确率(Average Accuracy)是 0.62，基于语料库方法的平均准确率为 0.63，基于语言学方法的平均准确率为 0.65。当本文使用的相似度阈值为 0.6 且间接连接概念节点权重为 0.6 时，本文准确度高达 69%。F-度量值是精度和召回率的平均值。当准确度为 69%时，本文的 F-度量值也达到最高 0.80。

实验表明，与其它相关方法相比，本文方法得到了较好的精度和召回率值，且本方法不需要从语料库获得任何外部知识。与基于词语相似度为基础的句子相似度计算方法相比，本文方法效率更高。总之，本文方法允许用户不需要任何额外的大文本语料库，仅基于领域本体结构就能进行高效的句子间相似度计算。这在构造新的应用时将是一个非常重要的特点。

参 考 文 献

[1] Burke R D, Hammond K J. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System[R]. TR-97-05. Univ. of Chicago, Dept. of Computer Science 1997

[2] Eugene A, Steve L, Luis G. Learning Search Engine Specific Query Transformations for Question Answering[C]// the Proceedings of the 10th World Wide Web Conference. 2001

[3] Ian G, Kai N Y. Eliminating Redundant and Less-Informative RSS News Articles Based on Word Similarity and a Fuzzy Equivalence Relation[C]// 18th IEEE International Conference on Tools with Artificial Intelligence. 2006: 465-473

[4] Ramiz M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization[J]. Expert Systems with Applications, 2009, 36(4): 7764-7772

[5] Mandreoli F, Martoglia R, Tiberio P. Searching Similar (Sub) Sentences for Example-Based Machine Translation[C]// Proceeding of 2002 Italian Symposium on Sistemi Evoluti per Basi di Dati. 2002

[6] Palakorn A, Hu Xiao-hua, Shen Xia-jiong. The Evaluation of Sentence Similarity Measures[C]// Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery. 2008: 305-316

[7] Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2008, 2(10)

[8] Jacob B, Benjamin C. Calculating the Jaccard Similarity Coefficient with Map Reduce for Entity Pairs in Wikipedia[OL]. http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf, 2008

[9] Metzler D, Bernstein Y, Croft W B, et al. Similarity measures for tracking information flow[C]// Proceedings of Information and Knowledge Management. 2005: 517-524

[10] Allan J, Bolivar A, Wade C. Retrieval and novelty detection at the sentence level[C]// Proceedings of SIGIR. 2003: 314-321

[11] Banerjee S, Pedersen T. Extended gloss overlap as a measure of semantic relatedness[C]// Proceedings of International Joint Conference on Artificial Intelligence. 2003: 805-810

[12] Ponzetto S P, Strube M. Knowledge Derived From Wikipedia for Computing Semantic Relatedness[J]. Journal of Artificial Intelligence Research, 2007, 30: 181-212

[13] Landauer T K, Foltz P W, Laham D. Introduction to Latent Semantic Analysis[J]. Discourse Processes, 1998, 25(2/3): 259-284

[14] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence[J]. Behavior Research Methods, Instruments & Computers, 1996, 28(2): 203-208

[15] Islam A, Inkpen D. Semantic similarity of short texts[C]// Recent Advances in Natural Language Processing. 2007: 227-231

[16] Li Y, McLean D, Bandar Z A, et al. Sentence similarity based on semantic nets and corpus statistics[J]. IEEE Transaction on knowledge and data engineering, 2006, 18(8): 1138-1150

[17] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity[C]// Proceedings of Association for the Advancement of Artificial Intelligence. 2006: 775-780

[18] Malik R, Subramaniam V, Kaushik S. Automatically Selecting Answer Templates to Respond to Customer Emails[C]// Proceedings of International Joint Conference on Artificial Intelligence. 2007: 1659-1664

[19] Chukfong H, Masrah A A M, Rabiah A K, et al. Word sense disambiguation based sentence similarity[C]// Proceedings of the 23rd International Conference on Computational Linguistics. 2010: 418-426

[20] Rejwanul H, Sudip K N, Andy W, et al. Sentence Similarity-Based Source Context Modelling in PBSMT[C]// 2010 International Conference on Asian Language Processing. 2010: 257-260

[21] Li Ru, Li Shuang-hong, Zhang Ze-zheng. The Semantic Computing Model of Sentence Similarity Based on Chinese FrameNet [C]// 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. 2009, 3: 255-258

[22] Li Yi, Liu Qiang. Chinese Sentence Similarity Based on Multi-feature Combination[C]// 2009 WRI Global Congress on Intelligent Systems. 2009,3:751-756

[23] Wang Rong-bo, Wang Xiao-hua, Chi zhe-ru, et al. Chinese Sentence Similarity Measure Based on Words and Structure Information[C]// 2008 International Conference on Advanced Language Processing and Web Information Technology. 2008:27-31

[24] Masrah H C F, Murad M A A, Doraisamy S C, et al. Measuring Sentence Similarity from Both the Perspectives of Commonalities and Differences[C]// 22nd IEEE International Conference on Tools with Artificial Intelligence. 2010,1:318-322

[25] Li Lin, Zhou Yi-ming, Yuan Bo-qiu, et al. Sentence Similarity Measurement Based on Shallow Parsing[C]// Sixth International Conference on Fuzzy Systems and Knowledge Discovery. 2009,7:487-491

[26] Shan Jian-fang, Liu Zong-tian, Zhou Wen. Sentence Similarity Measure Based on Events and Content Words[M]. Sixth International Conference on Fuzzy Systems and Knowledge Discovery,2009,7:623-627

[27] Lee Ming-che, Zhang Jia-wei, Lee Wen-xiang, et al. Sentence Similarity Computation Based on POS and Semantic Nets[C]// 2009 Fifth International Joint Conference on INC, IMS and IDC. 2009:907-912

[28] Dolan B, Quirk C, Brockett C. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources[C]// Proceedings of the 20thInternational Conference on Computational Linguistics. 2004

[29] http://www.WordNet.org/

(上接第 220 页)

6.4 基图像的稀疏度

首先定义度量稀疏度的函数[6]为：

$$sp(x)=\frac{\sqrt{n}-(\sum|x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n}-1} \tag{22}$$

式中，n 是向量 x 的维度，$0\leqslant sp(x)\leqslant 1$，当且仅当 x 仅有一个非零元时，函数值为 1；当且仅当所有的元素相等时，函数值为 0。

图 5 和图 6 分别表示在数据库 ORL 上特征维数 r 取值 25、在数据库 CBCL 上特征维数 r 取值 49 时，用 NMF 算法、GNMF 算法、GNMFSC 算法对由训练图像构成的非负矩阵 V 进行矩阵分解得到的基图像，其中用式(22)度量稀疏度，把基矩阵 W 看作向量，计算稀疏度。

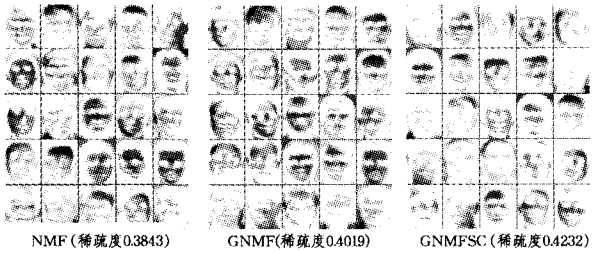


图 5　3 种算法在特征维数 r 为 25 时提取得到的基图像

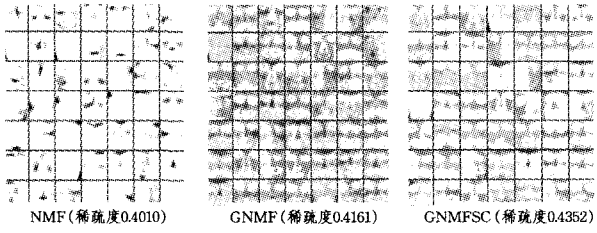


图 6　3 种算法在特征维数 r 为 49 时提取得到的基图像

由图 5 和图 6 可知，在这两个数据库上对比这 3 种算法的基图像稀疏度，NMF 的稀疏度最差，GNMFSC 的稀疏度最高，超过了 GNMF 基图像的稀疏度。也就是说，与 NMF 算法、GNMF 算法相比，GNMFSC 算法得到了最佳的局部表示。

结束语　提出了稀疏约束的图正则非负矩阵分解，并给出了迭代公式以及收敛性证明。以 ORL 人脸库和 CBCL 人脸库的数据作为实验对象进行验证，得出新算法具有更好的识别率，并提取到既稀疏又具有很强判别能力的基图像。

参 考 文 献

[1] 姜伟，杨炳儒. 局部敏感非负矩阵分解[J]. 计算机科学，2010,37(12):211-214

[2] Lee D D, Seung H S. Learning the parts of objects by non-negativematrix factorization[J]. Nature, 1999,401(6755):788-791

[3] Hoyer P O. Non-negative sparse coding[C]// Processings IEEE Workshop on Neural Netw. Signal Process. 2002:557-565

[4] Li S Z, Hou Xin-wen, Zhang Hong-jiang, et al. Learning spatially localized, parts-based representation[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2001,1:207-212

[5] Liu Wei-xiang, Zheng Nan-ning, Lu Xiao-feng. Nonnegative matrix factorization for visual coding[C]// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003,3:293-296

[6] Hoyer P O. Non-negative matrix factorization with sparseness cons-traints [J]. Journal of Machine Learning Research, 2004,5(9):1457-1469

[7] Wang Yuan, Jia Yun-de, Hu Chang-bo, et al. Fisher non-negative matrix factorization for learning local features[C]// The Fifth Conference on Computer Vision. 2004

[8] Zafeiriou S, Tefas A, Buciu I, et al. Exploiting discriminant information to frontal face verification [J]. IEEE Transactions on Neural Networks, 2006,17(3):683-695

[9] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computations, 2003, 15(6):1373-1396

[10] Cai Deng, He Xiao-fei, Jia Wei-han. Spectral regression: A unified approach for sparse subspace learning[C]// IEEE International Conference on Data Mining (ICDM). 2007

[11] Cai Deng, He Xiao-fei, Wu Xiao-yun, et al. Non-negative Matrix Factorization on Manifold[C]// The IEEE International Conference on Data Mining. 2008

[12] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]// Annual Conference on Neural Information Processing Systems. 2001,13:556-562