基于语义的文本流形研究

杨 震[1],范科峰[2,3],雷建军[4],郭 军[5]

(1. 北京工业大学计算机学院,北京 100124;2. 北京邮电大学网络与交换国家重点实验室信息安全中心,北京 100876;3. 中国电子技术标准化研究所,北京 100007;4. 天津大学电子信息工程学院,天津 300072;5. 北京邮电大学信息与通信工程学院,北京 100876)

摘 要: 本文通过引入包括 Isomap 流形降维、查询语义词典(WordNet)等高度非线性的方法,期望将文本信息处理领域长期专注于"语法"层次的研究,演进到"语义"的层次. 利用流形学习工具研究了中文词汇在语义空间(分类空间)的分布聚集情况,通过利用 WordNet 词典进行了短信聚类的研究. 实验结果表明,本文的方法能够更好地反映文本之间的内在联系.

关键词: 语义距离; 流形学习; 词汇分布; 短信聚类

中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0372-2112 (2009) 03-0557-05

Text Manifold Based on Semantic Analysis

YANG Zhen[1], FAN Ke-feng[2,3], LEI Jian-jun[4], GUO Jun[5]

(1. *School of Computer, Beijing University of Technology, Beijing* 100124, *China*; 2. *Information Security Center, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*; 3. *China Electronics Standardization Institute, Beijing* 100007, *China*; 4. *School of Electronic Information Engineering, Tianjin University, Tianjin* 300072, *China*; 5. *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: By using the non-linear operators (such as Isomap, WordNet etc.), how to promote the traditional text information processing techniques to "understanding" level was discussed. Based on manifold analysis, the distribution of Chinese words in a continuous semantic space was primarily studied. Short messages clustering based on WordNet was investigated. Experimental results prove that these methods can reflect the internal relation of texts.

Key words: semantic distance; manifold learning; words categorization; short message clustering

1 引言

文本信息处理的目的是使机器能够在"一定程度上理解并自动处理"文本信息. 换言之,技术只是手段,理解并自动处理才是目的. 但是我们首先面临一个问题:使机器能够在"一定程度上理解并自动处理"文本信息的基础是什么?是语素、词汇、词组或短语结构?是动名形副介连叹?是形形色色的逻辑语法?还是大规模真实语料及其统计结果?众多研究者给出了不同的答案. 但是句法-语义分析系统局限于小规模受限语言处理,无法胜任大规模真实文本信息处理的重任. 因此,当前各语种文本信息处理的主流技术仍是语料库方法和统计语言模型.

基于真实语料库的统计方法进行文本信息处理高度依赖于对数据模型的分析和可视化,在此过程中无法避免要处理大量的高维数据,即会遇到"维数灾难". 因此研究者经常会面临维数约简的问题:即如何发现高维数据的低维紧致表达. 从几何的观点来看,维数约简可以看成是挖掘嵌入在高维数据中的低维线性或非线性流形. 早在 20 世纪 80 年代末期,在 PAMI(IEEE Trans. on Pattern Analysis and Machine Intelligence)上就已经有流形模式识别的说法. 1995 年,"流形学习"(Manifold Learning)这一术语在语音理解的文献[1]中被首次引入. 2000 年,美国《Science》上发表 3 篇论文[2~4],从认知上讨论了流形学习,并正式使用了 Manifold Learning 的术语,强调认知过程的整体性. 从此,流形学习的过程变得有章

收稿日期:2008-03-20;修回日期:2008-12-10

基金项目:国家重点基础研究发展计划资助("973"计划)(No. 2007CB311100, No. 2007CB311203);国家自然科学基金(No. 60672112, No. 60702031);国家自然科学基金委员会——广东联合基金重点资助项目(No. U0835004);北京市教育委员会科技发展计划面上项目(No. KM200810005030);高等学校博士学科点专项科研基金(No. 20070013007);高等学校学科创新引智计划资助项目(No. B08004);北京工业大学博士科研启动基金(No. 52007999200703)

可循:首先对嵌入映射或者低维流形进行某种特定的假设,或者以保持高维数据的某种性质不变为目标,然后将问题转化为求解优化问题并且提供有效算法的支持.近年来,流形学习领域产生了大量的研究成果,包括LLE[2]、Isomap[4]、Hessian LLE[5]、Laplacian Eigenmap[6]、Diffusion Map[7]是几种有代表性的流形学习方法.

在文本信息处理领域,流形学习方法也得到了越来越广泛的应用.值得一提的有 Roweis S 和 Saul L[2]将 LLE 方法应用于发现英文词汇在语义空间中分布的尝试,Kouropteva O 等人基于 LLE 和 SVM 所做的关于手写数字识别的研究[8].这些研究成果都体现了流形学习方法在文本信息处理领域的成功应用.

本文研究的切入点是以统计语言学的角度,即从语料库语言学的角度,通过引入非线性操作,包括 Isomap 流形降维、查询语义词典(WordNet)等高度非线性的方法,期望将长期专注于"语法"层次的研究,演进到"语义"的层次.首先,研究了中文词汇在语义空间(分类空间)的分布聚集情况.其次,在此基础上,本文通过利用 WordNet 词典进行了短信聚类的研究.实验结果表明,本文的方法能够更好地反映文本之间的内在联系.

2 中文词汇在语义空间中的分布

本节研究了中文词汇在语义空间(分类空间)中的分布情况.这里需要说明的是,这里的语义空间的定义和传统语言学上"语义"的定义有所不同,特指文本分类空间.我们用语义空间这样的表述是为了说明,所做的研究已经脱离的原先语法的层次,向更易于为人所理解的层次进了一步.通过流形学习的方法,研究中文词汇在语义空间(分类空间)中的二维和三维的流形嵌入情况.值得注意的是,中文词汇在嵌入空间的坐标是有意义属性的,例如不同的距离代表着词汇之间不同的语义联系(semantic associations).类似的工作包括 Roweis S 和 Saul L[2]在 Lee D 和 Seung H[9]工作的基础上,研究了在 Grolier's Encyclopedia 中 31,000 篇文档中 5000 个词在这些文档空间中的流形结构.他们通过计算每个词的文档词频向量(word-document vector)之间的点积来表示相似性,然后使用 LLE 算法发现这些词在文档空间的低维流形嵌入,进而揭示出词在语义空间中的近似程度.

在本节中,借由 2004 年国家 863 文本分类评测语料库,使用 Isomap[4]探索中文词汇在文本分类空间中的低维流形,进而从分类意义上研究中文词汇之间的相似性,我们认为这种相似性也从语义层面上展示了中文词汇之间的相似关系.具体处理步骤如下:

(1)首先,确定合理的类别体系.本节采用的是《中国图书馆图书分类法(第四版)》(简称中图分类法),其中由于"T-工业技术"和"Z-综合性图书"这两个类别难以判定,不予考虑.因此类别体系中包含 21 个基本大类,加上 T 类工业技术中的 15 个二级类别,总共形成了 36 个类别.实验中采用的测试语料来源于 2004 年国家 863 文本分类评测所提供的测试库,其中提供了 3600 个不同的训练样本,但其中有 88 个分属两类,所以可将其看成 3688 个样本.

(2)其次,对文本库中的文本进行分词处理.文中采用的是使用 POC-NLW 语言标记模板[10]的分词方法.

(3)再次,对于分词后产生的词表 $W=\{w_1,w_2,w_3,\cdots,w_n\}$ 进行处理.先去除分词程序所产生的碎片,包括过长、过短以及无意义的片断等.然后将每一个词 w_i, $i=1,2,3,\cdots,n$,用其<文档频率-类别>向量表示,即用其对应的 36 个类空间(用 $A,B,C,\cdots,X$ 对应中图分类法中相应的类别)中的文档频率(*DF*:Document Frequency)来表示.对于词 w_i 将其在每个类别空间中 *DF* 组成一个向量:

$$\mathrm{DF}(w_i)=\{DF^A_{w_i},DF^B_{w_i},DF^C_{w_i},\cdots,DF^X_{w_i}\} \tag{1}$$

其中 $DF^j_{w_i}, j\in\{A,B,C,\cdots,X\}$ 表示 w_i 在标记为 j 类的文档中的 *DF* 值.将 $DF^j_{w_i}$ 用其各自类别的文档总数进行归一化处理,得到归一化的<文档频率-类别>向量:

$$|\mathrm{DF}(w_i)|=\{\frac{DF^A_{w_i}}{N^A},\frac{DF^B_{w_i}}{N^B},\frac{DF^C_{w_i}}{N^C},\cdots,\frac{DF^X_{w_i}}{N^X}\} \tag{2}$$

其中 $N^k, k\in\{A,B,C,\cdots,X\}$ 表示类别 k 中的文档总数.

这样一来,两个词 w_i 和 w_j 之间的距离就可以用其<文档频率-类别>向量之间距离表示.当然,距离的计算可以有很多种测度,如 *K-L* 距离、Hamming 距离、街区距离等,在这里使用的是平方距离:

$$Distance(w_i,w_j)=\sqrt{(\frac{DF^A_{w_i}}{N^A}-\frac{DF^A_{w_j}}{N^A})^2+\cdots+(\frac{DF^X_{w_i}}{N^X}-\frac{DF^X_{w_j}}{N^X})^2} \tag{3}$$

(4)由此,可以得到了词汇之间的两两距离(pairwise distance),进而可以计算 $W=\{w_1,w_2,w_3,\cdots,w_n\}$ 所包含词之间的距离矩阵:

$$D=\begin{Bmatrix} Distance(w_1,w_1) & \cdots & Distance(w_1,w_n)\\ Distance(w_2,w_1) & \ddots & \cdots\\ \cdots & \ddots & \cdots\\ Distance(w_n,w_1) & \ddots & Distance(w_n,w_n)\end{Bmatrix}_{n\times n} \tag{4}$$

当然,考虑到矩阵的对称性,可以只计算下三角矩阵.

(5)最后,得到 D 后,通过使用 Isomap 方法,就可以得到中文词汇 $W=\{w_1,w_2,w_3,\cdots,w_n\}$ 在分类空间中的分布情况.

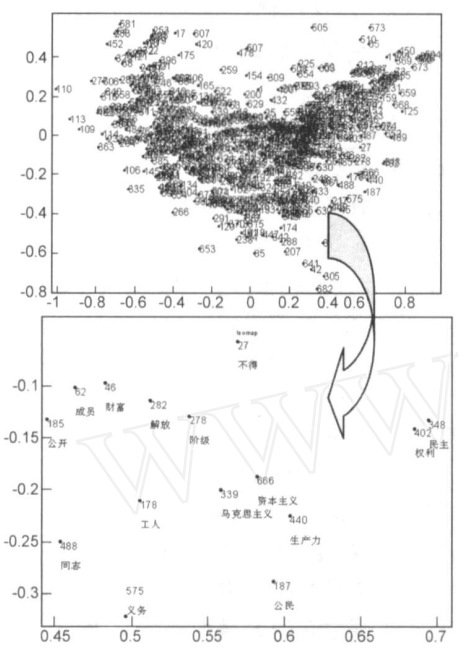

图1　中文词汇在文本分类空间（语义空间）的排列情况及其局部放大图。

实验中选择在中图分类法 36 个类中 *DF* 总和在[120,200]区间内的词条进行可视化. 其中 Isomap 算法邻居数 $k=6$. 中文词汇在分类空间(语义空间)中二维分布如图 1 所示. 选择降维到二维的情况是出于可视化的考虑,即使词的邻近不仅表征语义上的近似关系,而且这种近似关系还具有视觉直观. 从图中可以看出,中文词汇的确在分类空间(语义空间)具有低维流形嵌入,从图 1 的局部放大图可以看到,在分类空间(语义空间)中语义相近的词是聚集在一起的,如图中所示,政治类的词汇包括阶级,解放,马克思主义,生成力等等词汇是聚集在一起的. 实验的成功将为进一步进行基于语义的特征选择打下良好的基础. 当然首先必须克服的困难就是降低 Isomap 算法的计算复杂度.

3　基于 WordNet 的短信聚类

现今短信在舆论导向和传播上扮演着越来越重要的角色,短信已被一些学者誉为继报纸、广播、电视、网络之后的第 5 大媒体. 短信作为全新的文本媒体对象,具有其自身特点:

(1) 长度比较短,一般来说不超过 140 个字符;(2) 短信用语有其自身的特点,例如频繁使用网络语言以及缩略语;(3) 短信库存在大量重复. 由于群发,转发,下载短信行为频繁,短信库中存在大量的重复短信.

这使得基于短信的处理必须针对其自身的特点进行. 一般来说,文本特征项可以是字、词、短语、词对、段落、文摘甚至是篇章结构. 在目前的文本处理的应用中,优缺点忽现. 但是针对短信对象,这些文本特征表示方式在一定程度上都丧失在处理文本文档、web 网页、邮件、Blog 等长文本时所具有的良好的表达能力. 这使得传统的聚类分析方法在短信表示这个层次上就遇到了极大的困难. 无论是用传统的文本表示模型,包括布尔模型,概率模型,向量空间模型,还是用现在一些新兴的文本表示模型,包括背景语言模型,词对模型,相似性模型,都无法良好的表示. 总会遇到特征向量稀疏性的问题,最终使得短信的聚类的变为简单层次上"词重现"一级的短信聚集. 毫无疑问,对概念的准确表达及正确使用将会对短信/短文本聚类处理带来很大的帮助. 而本文也正试图从概念的高度进行短信的聚类分析,我们做出以下假设:(1) 短信在语义空间存在低维流形嵌入;(2) 短信在嵌入空间的几何关系反映他们之间的语义联系,因此在低维流形嵌入上进行聚类分析更能反映短信之间的相似程度;(3) 将两条短信之间的相似程度定义为其所组成词的某种定义下的匹配程度;(4) 将词的近似程度和其在 WordNet(网址:http://wordnet.princeton.edu/)组织结构中所处的不同位置联系在一起. 具体处理步骤如下:(1) 预处理步骤. 对于短信 $SM_i, i=1,2,3,\cdots,k$,首先将其分散成 m 个词块(token)的列表,即 $SM_i=\{T_1,T_2,T_3,\cdots,T_m\}$,其中每个 T_i 表示一个词块. 显然,为了得到词块,对于中文短信而言,须进行分词处理,而对于英文短信则需要进行预处理(去停用词,还原词干等);(2) 对词块(token)进行词性消歧和词义消岐;(3) 计算各个词块(token)之间的相似程度,可以使用各种基于编辑距离的方法. 在本节中使用的是基于 WordNet 的相似度的计算方法:将 WordNet 组织结构当作一个无向图,这样两个词块(token) T_i, T_j 之间的距离 $Distanc(T_i,T_j)$ 可以定义为其在无向图中的距离. 正如 Resnik P[11]的研究所证明的如果两个词块(token)之间的距离越近,那么这两个词块(token)就越相似. 那么对两个词块(token) $T_i, T_j, i, j=1,2,3,\cdots,n$ 之间的相似度可以如下定义[11]:

$$Simility(T_i,T_j)=1/Distance(T_i,T_j) \tag{5}$$

(4) 在定义了词块(token) T_i, T_j 之间的相似度后,可以进一步计算短信之间的相似度了. 假设短信 SM_i, SM_j 分别由 m 个词块和 n 词块组成:分别由$\{T_1^{SM_i}, T_2^{SM_i}, T_3^{SM_i},\cdots,T_m^{SM_i}\}$和$\{T_1^{SM_j}, T_2^{SM_j}, T_3^{SM_j},\cdots,T_n^{SM_j}\}$表示. 那么短信 SM_i, SM_j 之间的相似度就可由其包含词块(token)之间的相似度计算而来.

首先把$\{T_1^{SM_i}, T_2^{SM_i}, T_3^{SM_i},\cdots,T_m^{SM_i}\}$和$\{T_1^{SM_j}, T_2^{SM_j}, T_3^{SM_j},\cdots,T_n^{SM_j}\}$, $i=1,2,3,\cdots,m, j=1,2,3,\cdots,n$ 之间的相似度矩阵定义为[14,15]:

$$\begin{pmatrix} Similty(T_1^{SM_i}, T_1^{SM_j}) & \cdots & Similty(T_1^{SM_i}, T_n^{SM_j}) \\ \cdots & \ddots & \cdots \\ \cdots & \ddots & \cdots \\ Similty(T_m^{SM_i}, T_1^{SM_j}) & \cdots & Similty(T_m^{SM_i}, T_n^{SM_j}) \end{pmatrix}_{m \times n} \quad (6)$$

然后,用 Hungarian 算法[12]找到$\{T_1^{SM_i}, T_2^{SM_i}, T_3^{SM_i}, \cdots, T_m^{SM_i}\}$和$\{T_1^{SM_j}, T_2^{SM_j}, T_3^{SM_j}, \cdots, T_n^{SM_j}\}$之间的最大匹配.即将其视为二部图的最大匹配问题,把词之间的相似度看成连接权重.设 SM_i 在 SM_j 中的最大匹配是$\{T_{j_1}^{SM_j}, T_{j_2}^{SM_j}, T_{j_3}^{SM_j}, \cdots, T_{j_m}^{SM_j}\}$, $j_k \in \{1,2,3,\cdots,n\}$, $k = 1,2,3,\cdots,m$. SM_j 在 SM_i 中的最大匹配是$\{T_{j_1}^{SM_i}, T_{j_2}^{SM_i}, T_{j_3}^{SM_i}, \cdots, T_{j_m}^{SM_i}\}$, $j_k \in \{1,2,3,\cdots,m\}$, $k = 1,2,3,\cdots,n$.那么,短信 SM_i, SM_j 之间的距离可以如下定义[11,13,14]:

$$Distance(SM_i, SM_j) = f(A,B) = (|A| + |B|)/A \cdot B \quad (7)$$

其中:

$A = \text{sum}\{similarity(T_1^{SM_i}, T_{j_1}^{SM_j}), \cdots, similarity(T_m^{SM_i}, T_{j_m}^{SM_j})\}$, $B = \text{sum}\{similarity(T_1^{SM_j}, T_{j_1}^{SM_i}), \cdots, similarity(T_n^{SM_j}, T_{j_n}^{SM_i})\}$, $|A|$和$|B|$分别为 SM_i 和 SM_j 的长度.

这样就可以计算两条短信之间的语义距离.这样的做法比较适用于所计算文本比较短的情况,因为无论 Hungarian 算法还是查 WordNet 词典都是计算复杂度相当高的操作.

(5)使用 Isomap 方法发现短信在语义空间中的嵌入情况;(6)然后在嵌入上进行短信聚类分析.

算法的工程实现中使用了大量开源代码:Porter 词干还原程序(网址:http://www.tartarus.org/~martin/PorterStemmer);Brill Tagger 词性消歧程序(网址:http://www.cs.jhu.edu/~brill/acadpubs.html);Lesk 词义消歧算法[15];WordNet.net 语义词典;Dao T 开发的 WordNet.net 的.net 程序接口和主程序框架[13,14].

本节的研究中使用了 IJCAI 的短信数据库(网址:http://research.ihost.com/and2007/index.html)为研究对象(此短信数据库一共 854 条短信,经由人工整理为标准格式并且在词一级基本进行对齐的).聚类中所使用的算法是 KMEANS,将其聚为 5 个类.在语义空间中的嵌入情况如图 2,图 3 所示.其中,A 类短信 58 条,B 类短信 165 条,C 类短信 88 条,D 类短信 430 条,E 类短信 113 条.从聚类结果分析来看,基于 WordNet 短信聚类还是可行的:

首先,由于短信自身的特点,使得传统的聚类分析方法在短信表示层次上遇到了极大的困难,无论是用传统的文本表示模型,还是用现在一些新兴的文本表示模型,都无法良好的表示.总会遇到特征向量稀疏性的问题,最终使得短信的聚类的变为简单层次上"词重现"一级的短信聚集.而本节使用的方法,绕开了文本表示的问题,通过直接计算短信词块(token)之间的相似度的办法计算短信之间的相似性,最后利用 Isomap 发现其内蕴的流形结构,再在此流形结构上进行聚类分析,因而能够取得比传统方法优异的结果.

其次,和以前基于词频统计的聚类方法类似,本节的方法能保证形式上相似或一致的短信(如 E 类短信中有三条重复短信:短信库中的原始标号为 33,47,56;A 类短信中有两条重复短信:短信库中的原始标号 29,43)聚在一起(映射为一点).

再次,基于语义的短信聚类能够将内容意思相似的短信聚集在一起,这也是其它聚类算法不足之处,如 B 类短信基本上都是情感交流类的短信.

最后,由于短信库自身的限制:数量比较少,而且来源单一,这为聚类分析带来了困难.从聚类的结果来看,所有聚类类别之间的差异性并不显著.在我们的方法中,短信之间的相似度可以由其邻近程度来表示.这样能够很容易的找到和某一条短信最相似的 n 条短信,只要通过简单的邻域搜索就可以确定.这将为进一步基于内容的检索工作打下的基础.

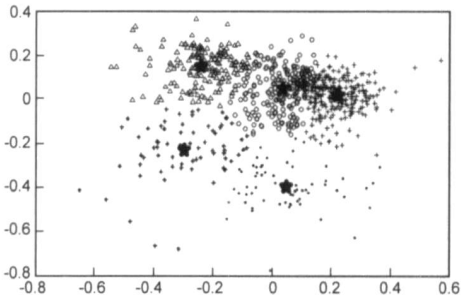

图2 短信在连续语义空间的排列情况(二维):(a)'*'A类短信;(b)'o'B类短信;(c)'+'C类短信;(d)'Δ'D类短信;(e)'.'E类短信.其中★代表每类的中心

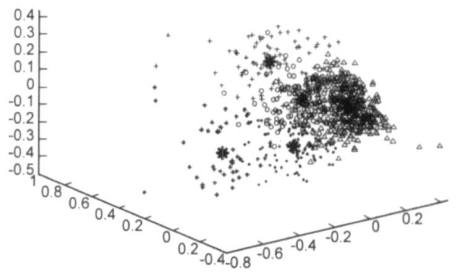

图3 短信在连续语义空间的排列情况(三维):(a)'*'A类短信;(b)'o'B类短信;(c)'+'C类短信;(d)'Δ'D类短信;(e)'.'E类短信.其中*代表每类的中心

4 结论

本文首先通过使用流形学习工具,研究了中文词汇在语义空间(分类空间)的分布情况.在此基础上,利

用 WordNet 词典进行了短信聚类的分析.研究结果将为进一步的基于语义的特征选择和信息检索工作打下基础.需要指出的是,本文的工作还比较初步,有待深入,至少在两个方面需要进一步的研究:

(1) 寻找更有效的短信相似度计算方法.如能更加充分地利用 WordNet 所能提供的信息,包括同义词词集(Synset)、类属信息(Class)、意义解释(Sense explanation),来共同来计算词的相似程度,相信可以更好的计算和模拟文本之间的相似程度;

(2) 寻找更有效的流形学习算法.现有的算法计算复杂度较高,而且缺乏效率良好的显式映射(explicit mapping),这些都在不同程度上制约了算法的实际应用.因此,开发更有效的有显式映射的流形学习算法是我们下一步的研究方向.

参考文献:

[1] Bregler C, Omohundro S. Nonlinear manifold learning for visual speech recognition[A]. Proc of Fifth Int. Conf. on Computer Vision[C]. Washington, DC, USA: IEEE Computer Society, 1995. 494.

[2] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323 - 2326.

[3] Seung H S, Lee D D. The manifold ways of perception[J]. Science, 2000, 290(5500): 2268 - 2269.

[4] Tenenbaum J, Silva D D, Langford J. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319 - 2323.

[5] Donoho D, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for highdimensional data[J]. PNAS, 2003, 100(10): 5591 - 5596.

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

[7] Coifman R, Lafon S, Lee A, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusionmaps[J]. PNAS, 2005, 102(21): 7426 - 7431.

[8] Kouropteva O, Okun O, Pietikäinen M. Classification of handwritten digits using supervised locally linear embedding algorithm and support vector machine[A]. Proc of the 11th European Symposium on Artificial Neural Networks[C]. Bruges, Belgium: D-side publi, 2003. 229 - 234.

[9] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401: 788 - 791.

[10] Chen B, He H, Xu W, et al. POC-NLW template based tagging method for Chinese word segmentation[A]. Proc. 2006 Int. Conf. on Computational Intelligence and Security[C]. Guangzhou, China: IEEE, 2006. 1423 - 1428.

[11] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language[J]. Journal of Artificial Intelligence Research, 1999, 11: 95 - 130.

[12] Alsuwaiyel M. Algorithms Design Techniques and Analysis[M]. Beijing: Publishing House of Electronics Industry, 2003.

[13] Dao T. An improvement on capturing similarity between strings[EB/OL]. http://www.codeproject.com/cs/algorithms/improvestringsimilarity.asp, 2005-08-05.

[14] Dao T. WordNet-based semantic similarity measurement[EB/OL]. http://www.codeproject.com/cs/library/semanticsimilaritywordnet.asp, 2005-10-01.

[15] Lesk M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine code from an ice cream cone[A]. Proc of the 5th Annual Int. Conf. on Systems Documentation[C]. Toronto, Canada: ACM, 1986. 24 - 26.

作者简介:

杨　震　男,博士,1979 年生于贵州六盘水,北京工业大学讲师.主要研究方向为信号处理、内容安全、可信计算.

E-mail: yangzhen@bjut.edu.cn

范科峰　男,博士,1978 年生于陕西礼泉,中国电子学会高级会员.主要研究方向为数字版权管理、无线通信、信号处理等.

E-mail: fankf@ccsi.ac.cn