基于跨语言广义向量空间模型的跨语言文档聚类方法

唐国瑜[1]，夏云庆[1]，张 民[2]，郑 方[1]
[1]清华大学 计算机科学与技术系，北京 100084
[2]资讯通信研究院，新加坡
E-mail: sweetyuer@gmail.com; yqxia@tsinghua.edu.cn; fzheng@tsinghua.edu.cn; mzhang@i2r.a-star.edu.sg

摘 要：跨语言文档聚类主要是将跨语言文档按照内容或者话题组织为不同的类簇。本文通过采用跨语言词相似度计算将单语广义向量空间模型（Generalized Vector Space Model, GVSM）拓展到跨语言文档表示中，即跨语言广义空间向量模型（CLGVSM），并且比较了不同相似度的在文档聚类下的性能。同时提出了适用于 GVSM 的特征选择算法。实验证明，采用 SOCPMI 词汇相似度度量算法构造 GVSM 时，跨语言文档聚类的性能优于 LSA。

关键词：跨语言文档聚类；跨语言广义向量空间模型；文档聚类；跨语言信息检索

Cross-lingual Document Clustering Based on Similarity Space Model

Tang Guoyu[1], Xia Yunqing[1], Zhang Min[2], Thomas Fang Zheng[1]
[1] Department of Computer Science and Technology,Tsinghua University, Beijing 100084
[2] Institute for Infocomm Research, A-STAR, Singapore
E-mail: sweetyuer@gmail.com; yqxia@tsinghua.edu.cn; fzheng@tsinghua.edu.cn; mzhang@i2r.a-star.edu.sg

Abstract: Cross-lingual Document Clustering is the task to automatically organize a large collection of cross-lingual documents into groups according to their contents or topics. This work extends traditional monolingual Generalized Vector Space Model (GVSM) to Cross-lingual GVSM (CLGVSM) by using cross-lingual term similarity calculation methods in order to represent documents in different languages and compare different term similarity calculation methods in cross-lingual document clustering. This work also proposes new feature selection method for CLGVSM. Experiment results show that GVSM with Second Order Co-occurrence Pointwise Mutual Information (SOCPMI) term similarity measure outperforms the latent semantic analysis (LSA) method.

Keywords: cross-lingual document clustering; text similarity; document clustering

1 前言

文档聚类的目的是按照相似程度将文档划分为不同的类簇，它已经成功应用于改进文档分类和文档流事件发现。国内外学者在单语言文档聚类研究中尝试了很多算法。但是随着商业环境的全球化，文档聚类逐步面临不同语言的挑战。

传统单语言文档聚类方法采取向量空间模型（Vector Space Model, VSM）表示文本，它利用词袋（Bag of Word, BOW）模型来构建特征空间，将每个文档转化为一个向量。词袋模型在特征匹配中通常采用“硬匹配”方法。例如，当词“海岸”被选为特征时，除非“海边”也被选为特征，否则“海边”无法影响到文档表示。这是因为“硬匹配”中“海岸”和“海边”完全不同。为解决这个问题，文献[1]提出的潜语义分析（LSA）方法，基于语料库中的共现信息将一组词与一个特征进行匹配。GVMS 则将文档中的词看做向量，然后通过计算词的内积或者相似度进行将文档表示在一个非正交的空间上。但是上述模型都是基于单语文档集设计的，并不能直接用到跨语言文档集中。

研究者提出了用词典或机器翻译工具对特征或者文档进行翻译。然而，一旦词被选为特征，“硬匹配”问题变得更为严重。如何对不同语言文档中的相似词汇，这是跨语言文档聚类的核心问题。文献[2]提出了采用 LSA 的解决方法。借助平行语料，他们将相似的词看作为一个特征。与单

语言 LSA 不同，跨语言 LSA 在固定训练集上选择特征。但由于目标文档集通常与训练集存在内容和用词的显著不同，这会导致过度适应问题。

本文通过采用跨语言词汇相似度计算将单语广义向量空间模型（Generalized Vector Space Model, GVSM）拓展到跨语言文档表示中，即跨语言广义空间向量模型（CLGVSM）。同时提出了适用于 CLGVSM 的特征选择算法。本文实现了两种有代表性的词汇相似度算法，即基于知网的词汇相似度算法和基于 SOCPMI 的词汇相似度算法。实验表明，SOCPMI 比知网更适合文档聚类。同时，我们还在相同可比语料下对基于 SOCPMI 的 CLGVSM 方法与 LSA 方法进行了对比。实验结果表明，基于 SOCPMI 的 CLGVSM 方法比 LSA 方法显示出更好的性能。

2 相关工作

2.1 跨语言文档聚类

跨语言文档聚类的难点在于如何处理跨语言相似性问题，其中最直接的方法是采用词典或机器翻译工具。在 TDT-3 评测中，四个系统均采用机器翻译工具（文献[3]等）。结果表明，与单语言话题跟踪相比，采用机器翻译方法会导致 50%的性能下降。下降的主要原因是机器翻译技术的准确性问题。

一些研究工作[4][5][6]通过双语词典进行词匹配或者特征词翻译。文献[7]通过多语言主题词表 Eurovoc 构造跨语言文档向量。以上基于词典的跨语言文档聚类方法都难以解决歧义词的翻译问题以及未登陆词问题。

近年来，学者开始利用平行语料或可比语料进行跨语言文档聚类[2][8]。不同于文档分类，文档聚类缺乏训练数据，因此语义空间只能在固定训练语料中构建，特征的选择也是如此，因此忽略了特征在聚类目标集中的不同分布。本文提出的 CLGVSM 模型构建于词汇相似度之上并在聚类目标集中进行特征选择。

2.2 词汇相似度

词汇相似度计算是一个自然语言处理研究热点，并在机器翻译和词义排歧等研究中得到应用。近年来提出的词汇相似度计算算法或基于统计技术，或基于语义网络。文献[9][10]提出基于 WordNet 的英文语义相似度计算方法。文献[11][12][13]则提出了利用知网概念定义计算跨语言词汇相似度的方法。基于语料的词汇相似度计算方法更为广泛。最经典的方法是点互信息（Pointwise Mutual Information, PMI）[13]。PMI 值越大，说明词汇越有可能出现在同一语境下。文献[14]提出了基于 PMI-IR 的同义词获取方法，利用 Alta Vista Adanvced 搜索引擎计算单词之间的概率。LSA 方法[15]分析大规模语料，利用词汇之间的共现信息计算词汇和文本的相似度。SOCPMI 方法[16]利用 PMI 将两个目标词的相邻词按重要性排序，并通过计算相邻词的 PMI 实现目标词之间的相似度计算。

本文采用两个最具代表性的词汇相似度计算方法构造 CLGVSM 矩阵：基于知网的词汇相似度[12]和基于 SOCPMI 的词汇相关度[16]。

3 相似度空间模型

为了便于描述，我们首先介绍传统的广义向量空间模型。

3.1 广义向量空间模型

假设$D = \{d_j; j = 1, ...n\}$表示包含n个文档m个词的文档集。X表示一个$m \times n$的矩阵，它

的元素x_{ij}表示词t_i在文档d_j的权重。GVSM[17]将文档表示在一个非正交空间中，文档的相似度计算公式如下：

$$Sim^{\mathrm{GVSM}}(d_1,d_2)=\frac{d_1^TGd_2}{\sqrt{d_1^TGd_1}\sqrt{d_2^TGd_2}} \tag{1}$$

其中G是一个$m\times m$关联矩阵，用来表示词之间的相似度。

传统的GVSM中[18]，词表示为文档的对偶空间中的向量。G的计算公式如下：

$$G=XX^T \tag{2}$$

在改进的GVSM中[21]，性能最好的G为词向量的协方差矩阵。

$$G_{\mathrm{COV}}=\frac{1}{n_c-1}QHQ^T \tag{3}$$

其中Q为X的抽样，并且$H=1-\frac{1}{n_c}ee^T$。

在上述GVSM模型中，G都是在聚类文档中计算得出的，但是它们很难获得跨语言的词信息。因此我们通过采用跨语言词相似度计算将GVSM拓展为跨语言文档表示模型CLGVSM。

3.2 跨语言广义空间向量模型上的特征选择和文档表示

VSM模型中，词对于一个文档的重要性可以简单采取词频表示，对于一个文档集的重要性则用倒文档频表示。拓展到CLGVSM模型中，我们定义了类似的特征重要性指标。

考虑一个包含“criminal”3次、“imprisonment”10次的文档。认为词“criminal”仍然是非常重要的，虽然他的词频比较低。这是由于“imprisonment”与“criminal”是语义相似的。为此，我们提出了两个基于CLGVSM模型的特征重要性指标:软词频和软文档频。给定词汇t和文档集D=$\{d_j\}_{j=1...L}$，假设d_j=$\{w_{i,j}\}_{i=1...N}$代表文档d_j的中的词汇，软词频和软文档频的定义如下：

软词频TF^s：

$$TF^s(t,d)=Sim^{\mathrm{SSM}}(v_t,d) \tag{4}$$

软文档频DF^s：

$$DF^s(t)=\textstyle\sum_{d_j\in D}\max_i Sim^{\mathrm{WD}}(t,w_{i,j})) \tag{5}$$

参考TF-IDF公式的思想，我们定义软倒文档频：

$$IDF^s(t)=log\Big(\frac{L}{DF^s(t)}\Big) \tag{6}$$

因此，词汇t在文档d的权重计算公式：

$$w^s(t,d)=TF^s(t,d)IDF^s(t) \tag{7}$$

如果我们单纯依靠权重进行特征选择，相似度比较高的单词会同时被选为特征。这是因为相似度比较高的单词含有相近的权重，这将造成特征集的冗余。因此，我们提出了一个改进的特征选择算法，只赋予相似词集中的一个词比较高的软词频，而其余词汇则降低权重。即按照初始软词频的从大到小更新软词频，删除相似度所造成的冗余。

对软词频改进后，我们根据公式（7）计算每个特征的权重，并按照特征权重的大小选择每个文档的特征，然后合并为一个特征集。我们使用特征集表示文档，并考虑特征集之外的词对文档表示的影响。我们将每个特征集外的词汇的软词频乘以相似度，累加到与它相似度最大的特征中，从而体现其贡献。这样，即使文档中并不包含某特征，文档表示也可以将文档映射到最有代表性的近义特征中。

3.3 基于广义空间向量模型的文档聚类算法

获得文档相似度后，我们采用聚类算法进行文档聚类。聚类算法不是本文的重点，因此我们选用经典的聚类算法，即 HAC（Hierarchical Agglomerative Clustering）算法[19]。

HAC 算法先将每个文档看成一个类簇，然后逐步将相似度最高的类簇合并为一个类簇。为了计算类簇之间的相似度，我们采用 group-average link 算法[19]。当类簇个数达到预定值后，则停止合并过程。

4 词汇相似度

词汇相似度在 CLGVSM 矩阵的构建中起到重要的作用。我们采用两种词汇相似度计算算法构造 CLGVSM 矩阵：基于知识的词汇相似度算法以及基于统计的词汇相似度算法。

文献[12]利用知网计算跨语言词汇相似度，基本思想是利用知网中词汇的语义定义。篇幅所限，详细过程参见文献[12]。

严格来说，基于统计的词汇相似度计算算法其实是与它们在语料中的共现程度有关。因此我们可以称统计的词汇相似度为词汇相关度。

由于 SOCPMI 在词汇相似度计算中具有优越性[16]，本文采取了这个算法。篇幅所限，详细过程参见文献[16]。

然而 SOCPMI 算法只能处理单语言的词汇相似度。本文扩展了这个算法，以实现跨语言词汇相似度计算。先在相同语言上对相邻词进行排序，然后计算它们的跨语言 PMI 值。

可以使用两种类型的语料计算跨语言词汇相似度：平行语料和可比语料。平行语料被广泛用于机器翻译，它是句子对齐的。但本文没有选用平行语料，原因有二：首先构造一个平行语料的成本比较高；其次跨语言的词汇相似度对句子对齐的要求并不高。最终本文选用更容易获得的篇章对齐的可比语料。

5 实验

5.1 实验设置

- 开发集

我们从英文和中文 GigaWord 中构建了一个中英文可比语料。我们采用以下的策略获得不同语言的可比文档对。1）文档相似度。采用基于 VSM 的文档相似度获得单语言中的可比文档。为了保证精度，我们设置文档相度的阈值为 0.4。2）基于知网获得词汇翻译。我们利用知网获得词汇之间的翻译信息，利用这些翻译信息计算跨语言文档那个相似度。3）时间限制。本文在计算文档相似度的时候还考虑到时间的限制，只选取在同一天内的新闻计算文档相似度获得可比语料。我们最后获得 101,409 篇中英文可比文档对。

- 测试集

我们采取 TDT4 数据集作为测试集。TDT4 数据集的信息如表 1 所示。

表 1　TDT4 数据集统计信息

语料	TDT41 (2002)	TDT42 (2003)
英文(话题数/文档数)	38/1270	33/617
中文(话题数/文档数)	37/657	32/560
总计(话题数/文档数)	40/1927	37/1177

- 评测指标

我们采用了文献[20]提出的评测指标。首先计算每个类簇最大的 F 值。假设A_i代表系统生成的类簇c_i的文档，A_j代表人工标注的类簇c_j的文档。则F值计算如下：

$$p_{i,j}=\frac{|A_i\cap A_j|}{|A_j|} \quad p_i=\max_j\{p_{i,j}\} \qquad r_{i,j}=\frac{|A_i\cap A_j|}{|A_i|} \quad r_i=\max_j\{r_{i,j}\}$$

$$f_{i,j}=\frac{2\cdot p_{i,j}\cdot r_{i,j}}{p_{i,j}+r_{i,j}} \quad f_i=\max_j\{f_{i,j}\}$$

其中$p_{i,j}$，$r_{i,j}$和$f_{i,j}$分别代表准确率、召回率和F值。

- 实验方法

本研究中，我们评测了以下五个方法：

VSM：采用VSM表示文档，并从知网获得词汇翻译信息。

LSA：LSA在可比语料中实现了文献[2]中的方法。

CLGVSM^HN：采用基于知网的跨语言相似度的GVSM。在GVSM矩阵的构造中，经过实验验证词汇相似度阈值为0.7。

CLGVSM^PMI：采用基于SOCPMI的跨语言相似度的GVSM。相似度阈值为0.4。

CLGVSM^PMI&TR：将SOCPMI与知网的翻译信息结合起来，知网获得翻译对的相似度为1。

5.2 实验结果及讨论

我们比较了五个系统在两个测试集上的性能。结果如表2所示。

表2 系统在两个测试集上的最高F值

方法:	VSM	LSA	GVSM^HN	GVSM^PMI	GVSM^PMI&IR
TDT41 最高F值	0.886	0.705	0.878	0.844	0.910
TDT42 最高F值	0.882	0.730	0.891	0.826	0.910

从表2可以得出如下结论：

首先，方法CLGVSM^HN 和 VSM 的性能相近，基于知网跨语言词汇相似度构造的GVSM比VSM几乎没有优势。观察发现，基于知网计算的相似度非常高。比如词“*Federal Reserve*”和“*bank*”的相似度为1。经过分析，基于知网的跨语言词汇相似度更多关注词的语义特征而不是语义本身，它倾向于给语义相似的词对更高的相似度，而不管它们是否是语义相关。这不利于文档聚类。因此可以认为，基于知网的词相似度不太适用于文档聚类。

其次，方法CLGVSM^PMI在两个测试集上的性能均优于方法LSA。在测试集TDT41上，F值提高了0.11。在测试集TDT42上F值提高了0.094。这说明了方法CLGVSM^PMI更适合跨语言文档聚类。分析原因如下：LSA 所构建的语义空间是在固定的可比语料中构建的，因此它没有考虑到目标聚类集的特征的重要性。相比之下，方法 CLGVSM^PMI 充分利用了测试集的信息构建语义空间。

最后，SOCPMI与知网相结合的CLGVSM的性能比较VSM的性能要好。在测试集TDT41中，高出0.014；而在测试集TDT42的效果更加明显，超出了0.018。这是本次实验获得最好结果(0.910)。这表明，使用恰当的词汇相似度计算方法，CLGVSM方法能取得满意的跨语言文档聚类效果。从表2可以看出，当只使用知网时，CLGVSM方法给出的结果与VSM相近。当只使用可比语料时，CLGVSM给出的结果比VSM要差。我们发现，从知网获得翻译信息非常重要。同时使用可比语料和知网，CLGVSM获得最好的性能。因此，知网与语料相结合可以获得更好的性能。

6 结语

本文的贡献主要有三个：（1）通过加入跨语言词汇相似度将 GVSM 拓展为 CLGVSM；（2）实现了基于知识和基于统计的词汇相似度计算方法。（3）对 CLVSMGVSM 方法和主要流行方法进行了评测，实验结果表明，利用知网以及可比语料资源，CLGVSM 模型比 VSM 和 LSA 的性能更优。

本文得出两个结论：首先，CLGVSM 方法比 VSM 和 LSA 都更有效；其次，结合知网翻译信息以及可比语料的相似度，有利于进一步提高文档聚类效果。在接下来的工作中，我们计划将 GVSM 模型用于更多语言的跨语言聚类。同时，由于 CLGVSM 模型能在语义空间上有效表示文本，我们将应用 CLGVSM 模型到短文本聚类中，希望能很好地解决稀疏问题。

参 考 文 献

[1] T. Landauer, P. W. Foltz and Darrell Laham. Introduction to Latent Semantic Analysis. Discourse Processes 25: 259-284.

[2] C-P. Wei, C. C. Yang and C-M. Lin. A Latent Semantic Indexing Based Approach to Multilingual Document Clustering. Decision Support System. 45(3): 606-620.

[3] T. Leek, H. Jin, S. Sista, and R. Schwartz. The BBN crosslingual topic detection and tracking system. Proc. of TDT'1999.

[4] H. H. Chen and C. J. Lin. A multilingual news summarizer. Proc. of COLING'2000: 159-165.

[5] D. K. Evans and J. L. Klavans. A Platform for Multilingual News Summarization, Technical Report. Department of Computer Science, Columbia University.

[6] B. Mathieu, R. Besancon and C. Fluhr. Multilingual Document Clusters Discovery. Proc. of RIAO'2004: 1-10.

[7] B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, I. Temnikova. Multilingual and cross-lingual news topic tracking. Proc. of COLING'2004: 959-965.

[8] D. Yogatamaand K.Tanaka. Multilingual Spectral Clustering Using Document Similarity Propagation. Proc. of EMNLP'2009: 871-879.

[9] D. Lin. Automatic retrieval and clustering of similar words. Proc. of COLING'98: 768-774.

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

[11] Q Liu, S Li. Word similarity computing basedonHowNet. Computational Linguistics andChinese Language Processing. (in Chinese)

[12] Y. Xia, T. Zhao, and P. Jin. Measuring Chinese-English Cross-lingual Word Similarity with HowNet and Parallel Corpus. Proc. of CICling'2011(II): 221-233.

[13] K.W. Church and P.Hanks. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1): 22-29.

[14] P. D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. Proc. of ECML'2001: 491-502.

[15] T. K. Landauer and S. T. Domais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquision, Induction and Representation of Knowledge. Psychological Review. 104(2): 211-240.

[16] A. Islam and D. Inkpen. Second order co-occurrence PMI for determining the semantic similarity of words. Proc. LREC'2006: 1033-1038.

[17] SKM. Wong, W. Ziarko, PCN. Wong. Generalized vector model in information retrieval. Proc. of the 8th ACM SIGIR: 18-25.

[18] A. K. Farahat, M. S. Kamel. Statistical semantic for enhancing document clustering. Knowledge and Information Systems.

[19] E. M. Voorhees. Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. Information Processing and Management, 22(6): 465-476.

[20] M. Steinbach, G. Kapypis, V. Kumar. A Comparison of Document Clustering Techniques.KDD Workshop on Text Mining, 2000: 109-111.