弱指导的统计隐含语义分析及其在跨语言信息检索中的应用*

金千里　赵军　徐波
中国科学院自动化所模式识别国家重点实验室
北京中关村东路 95 号，北京 2728 信箱，100080
{qljin, jzhao, bxu}@nlpr.ia.ac.cn

摘要：本文提出了一种语义聚类和扩展的新方法，称为有指导的统计隐含语义标引（SPLSI）算法。该算法能基于双语语料，通过机器学习来自动进行语义聚类，生成词间相似度矩阵。和以前的算法相比，SPLSI 算法不仅在聚类意义上更加明确、聚类的过程更容易控制，而且降低了时间和空间复杂度。基于 SPLSI 算法，实现了跨语言信息检索领域的三个系统：多语言文本分类，跨语言文本检索，跨语言关键词扩展。实验结果显示，在准确率、召回率、平均运算时间等多个评价指标中，SPLSI 均优于以前的各种算法。

关键词：隐含语义标引　跨语言信息检索　多语言文本分类 关键词扩展

Weakly-Supervised Probabilistic Latent Semantic Analysis and its Applications in Multilingual Information Retrieval

Qianli JIN, Jun ZHAO, Bo XU
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science
95 Zhong Guan Chun Dong Rd., Beijing, China, 100080
{qljin, jzhao, bxu}@nlpr.ia.ac.cn

Abstract: This paper proposes a new method for meaning clustering called 'Supervised Probabilistic Latent Semantic Indexing' (SPLSI). Based on bilingual corpora, the algorithm can produce words-similarity-matrix through machine learning. The advantage of SPLSI is that it is more reasonable and controllable in meaning clustering, but has less time complexity. And based on this method, we produce three application systems in the field of cross-lingual information retrieval (CLIR). They are 'multilingual text categorization, 'multilingual text retrieval' and 'key words expansion'. Experiments indicate that SPLSI outperforms the existing methods, and has good effectiveness in many applications of CLIR.

Key Words: latent semantic indexing (LSI), cross-lingual information retrieval (CLIR), multilingual text categorization

*本文受国家 973 项目子课题（G1998030501A-06）和国家自然科学基金项目（60272041）资助。

1 引言

1．1 跨语言信息检索的背景

随着互联网的普及，网上信息资源也越来越丰富。由此给信息检索（IR）带来两个问题，一是如何在 Internet 这样一个开放式的数据库中准确的找到相关信息，二是如何克服语言障碍（Language Barrier）问题，即实现跨语言的信息检索（CLIR）。

双语之间的跨语言信息检索，代表性的研究有：美国 Massachusetts 大学的 Lisa Ballesteros 和 W. Bruce Croft 的英语和西班牙语之间的交叉语言信息检索研究，采用的是双语词典结合译词选择排歧的方法；复旦大学吴立德和黄萱菁的英汉交叉语言信息检索研究；微软亚洲研究院高建峰等的英汉交叉语言信息检索研究；以及中国科学院软件研究所的英汉交叉语言信息检索研究； 这些研究工作主要都是基于双语词典和译词选择的方法，不能很好的解决语言障碍问题。美国 Duke 大学的 Michael L. Littman 将单语言信息检索中的隐含语义标引（LSI）扩展到双语信息检索中，形成 CL-LSI，试验取得令人满意的结果；但是由于 LSI 自身的物理意义不够明确，所以较难控制词义聚类的效果；此外这个算法的空间和时间复杂度太大，在目前的硬件条件下很难实际应用。1999 年，Hofmann 提出了统计隐含语义标引（PLSI）的概念，在理论和算法上都有所突破；目前还极少有这一技术在跨语言信息检索中应用的相关研究。

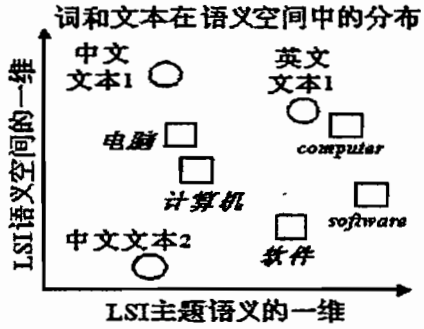
1．2 跨语言信息检索的算法分析

- 基于词典的方法。这种方法直接用词典进行全文翻译，类似于机器翻译的技术。这种方法代价太大，对于大文本集合是不可行的。并且，对于信息检索任务来说，对文本的完全翻译既是不必要的，也由于缺少上下文约束而使排歧很困难。
- 基于中间语言的方法。中间语言方法的一个主要优点是涉及到了双语之间的语义对应。这种方法实际上是第二种方法的一个延伸，只是把关键词替换为一种抽象的概念空间。但是，双语之间往往这种概念并不是很匹配的很好，尤其是对于两种不同风格的语言（如中文和英文）而言更是效果欠佳。
- 基于多语言对齐语料库的 **LSI** 方法。LSI 是“隐含语义标引”的简称，与上述方法不同的是，LSI 不再将词和文本之间的关系看成是孤立的，而是用一个相似度值来衡量。首先构造一个文档—词的相似度矩阵 X，矩阵中的每个元素是相应词在文档中出现的次数或者频度。根据矩阵分析中的奇异值分解（SVD）算法，得到：$X = U^T \sum V$ 。中间的对角阵中的元素，即奇异值。当把较小的奇异值忽略，可以大量的压缩空间，提高效率，此外还多了一个 smoothing 的过程。但是，SVD 算法速度太慢，且不具备物理意义，所以分类的聚合度无法控制。

1．3 本文的研究工作

基于以上分析，本文的工作主要有两个方面：第一是改进 PLSI 算法，求得对分类结果更好的控制，并且降低空间和时间复杂度，即称为“有指导的统计隐含语义标引”（SPLSI）；第二是把我们的 SPLSI 算法应用到跨语言信息检索中，以获得更好且更人性化的查询效果。

2 算法描述



SPLSI 是“有指导的统计隐含语义标引”的简称，与 LSI 方法不同的是，采用了概率模型来衡量文档—词之间的关系，并引入了隐含空间的概念。这样，每一文本就被映射到一个语义空间，每个词也同样。(见左图)。从而，无论是词与词之间还是文档与词之间，都通过这个语义空间来表达相似度，就有了相当明确的物理意义与可信度。在优化的决策函数方面，PLSI 采用的是最大熵模型，聚类效果要好于 LSI 的最小二乘模型。此外，PLSI 采用了迭代算法来实现，大大降低了时间复杂度，约为 LSI-SVD 的百分之一。

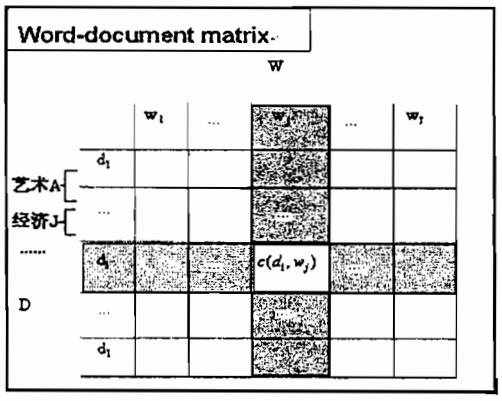
SPLSI 算法分为两部分：建立标引的算法和分类检索算法。前者是模型的建立过程，主要是为了求得一系列的概率相似度矩阵；后者是建立好的模型的应用，其中包括：文本检索（分类）算法，以及关键词扩展算法。下面将分别详细描述。

2．1 建立 SPLSI 标引的算法(基于中英文本对齐的双语语料库)

(a) 汉语分词与文本粗分类

首先采用是已经较为成熟的最大概率分词方法对中文文档分词。然后，需要对双语文本进行人工的粗分类，这个过程主要是为了对后续迭代算法的初值进行弱指导。从算法本身来讲，粗分类不是必须的。此处，我们简单的将双语文本分为以下的 10 类：艺术、经济、军事、政治、新闻、生活保健、体育、教育、法律、科学。将同类型的文本排列在一起，供下一步使用。由于是粗分类，所以不要求很高的准确率。

(b) 构造“文档—词”索引矩阵



如左图构造文档—词的索引矩阵 M(Word, Document)，其中的文档按照类型排序。矩阵 M 中元素的初始值 c(d,w)设为单词 w 在文档 d 中出现的次数。然后，需要进行归一化的操作，主要基于以下两个原因：第一，每篇文章中词的个数多少不同，因此一个词在短文章中出现一次的价值，显然应该大于在长文章中出现一次的价值；第二，一个很少出现的词，一旦出现在文档中，其价值应该大于普遍出现的词。事实上，类似于“the, 我们，的，of”之类的词几乎在任何文档中都会出现，因此其价值应该是趋向于零的。所以，归一化的工作有两个。首先是根据一个停用词表（stoplist），把 M 矩阵中所有停用词对应的列去除。然后根据公式（1）进行归一

化计算：$m(d,w)=\frac{c(d,w)}{Count\ (w)}\times\frac{\log\ \beta}{\log\ Length\ (d)}$..... (1)

其中，c(d,w)是矩阵 M 初始值，β 是系数，Count(w)是词 w 在所有文档中出现的总次数，Length(d)是文档 d 中所有非停用词数。

(c) 构造语义空间，确定映射初始值

构造 k 维的语义空间 Z，并且依据(a)中的粗分类结果给出语义空间的先验概率 p(z)。

具体的操作如下：设有 n 篇文档，文档共分为 t 种类型，其中第 1 篇到第 i 篇是同一类型的，那么有：$p(z_1)=p(z_2)=...=p(z_{[k/t]})=[\frac{i}{n}\times\frac{1}{[k/t]}]$ ..... (2)

其中，'[ ]' 表示取整操作，k 值的选取依赖于经验，如果太小则无法把各类分开，如果太大则太敏感，容易引入噪声；在一般应用中可取 20 到 100。

有了语义空间后，需要分别构造"文档—主题"的映射矩阵 P(D,Z)和"词—主题"的映射矩阵 P(W,Z)，并给出初始值。设共有文档 n 篇，其中文档 d 属于第一类，而第一类的文档共有 i 篇，则：(其余部分可以依次类推)

$$p(d,z_1)=...=p(d,z_{[k/t]})=\frac{1}{[k/t]};\text{其余}\ p(d,z)=0 \quad (3)$$

而对矩阵 P(W,Z)，由于不知道任何的先验知识，所以就给随机值作为其初始值；需要注意的是，必须满足概率矩阵的条件，也就是任何一行的值之和必须是 1。

(d) 采用 EM 迭代算法，求得结果

根据上述的结果，可以求得"文档—词"的相似度矩阵 P(W,D)初始值：

$$p(w,d)=\sum_{z\in Z}p(z\mid w)p(z)p(z\mid d) \quad (4)$$

然后，在最小熵的意义下，进行优化。即最大化以下函数（其中 m(w,d)是索引矩阵 M 中的元素）：$L=\sum_{w\in W}\sum_{d\in D}m(w,d)\log\ p(w,d)$ ..... (5)

对语义空间的每一维，采用标准的 EM 算法，进行迭代最优化。

E-Step: $p(z\mid w,d)=\frac{p(z)p(d\mid z)p(w\mid z)}{\sum_{z'\in Z}p(z')p(d\mid z')p(w\mid z')}$ ..... (6)

M-Step: $p(w\mid z)=\frac{\sum_{d\in D}m(w,d)p(z\mid w,d)}{\sum_{d\in D}\sum_{w'\in W}m(w',d)p(z\mid w',d)}$ $p(d\mid z)=\frac{\sum_{w\in W}m(w,d)p(z\mid w,d)}{\sum_{d'\in D}\sum_{w\in W}m(w,d')p(z\mid w,d')}$

$$p(z)=\frac{\sum_{d\in D}\sum_{w\in W}m(w,d)p(z\mid w,d)}{\sum_{d\in D}\sum_{w\in W}m(w,d)} \quad ..... \quad (7)$$

反复应用公式(6)(7)，直到函数(5)的变化量很小，即可认为达到了最大值。从而就获得了最优化的 *P(Z),P(W,Z),P(D,Z)*矩阵；再次利用公式 (4)，可以求得"文档—词"相似度矩

阵 P(W,D); 进一步利用（8）式，可以得到“词—词”相似度矩阵 P(W,W)。

$$p(w_1, w_2) = \sum_{d \in D} p(w_1 \mid d)\, p(d \mid w_2) \qquad \cdots\cdots \qquad (8)$$

在 P(W,D)和 P(W,W)矩阵中，中文词和英文词的地位是完全等同的。

（e）模型的动态扩展与优化

如果已经构造并且最优化了上述的一组矩阵，当有一些双语新文本 d'到来的时候，我们拥有了新的知识。如何将这些新知识融合到这组矩阵中就成为一个问题。虽然 EM 算法本身是无法打断的，但可以有一种变通的办法来处理。

如果我们已经有了 *P(W,Z),P(D,Z)*等一组矩阵，当文档 d'到来的时候，首先根据（b）中的步骤构造子 d'的标引矩阵 *M'(W,D')*，并把它拼接到原始的 *M* 中，形成一个新的 *M* 矩阵。这一步增加了文档的维度，而词的维度保持不变。然后，根据(c)中的步骤构造子矩阵 *P'(D',Z)*，同样的将其拼接到 *P(D,Z)*中。重新进行 EM 迭代，就可以扩展模型，加入新知识了。

2．2 SPLSI 算法的应用（基于 2.1 中的模型）

（a）SPLSI 多语言文本分类

文本分类问题的核心是计算文本之间相似度。设从文本 do 中抽取词向量 Wo，其维度等于 *P(W,W)*矩阵的行向量维度，其元素 Wo(word)为词 word 在文本中出现次数的归一化值。利用 *P*，得到文本相似度：

$$Similarity(do, dn) = Wo \bullet P(W, W) \bullet Wn^T \qquad \cdots\cdots \qquad (9)$$

（b）SPLSI 跨语言文本检索

和（a）中相似，跨语言文本检索的核心问题是关键词向量和文本的相似度计算。把用户查询的关键词构造成词向量 Wq，其维度等于 P(W,W)矩阵的行向量维度。则查询主体词和文本相似度为：

$$Similarity(Query, dn) = Wq \bullet P(W, W) \bullet Wo^T \qquad \cdots\cdots \qquad (10)$$

(c) SPLSI 跨语言查询关键词扩展

基于 SPLSI 的跨语言关键词扩展，实际上整合了机器翻译，词义消歧，语义扩展等多项功能。所有的工作综合起来，乘一个词间相似度矩阵即可完成。首先构造查询关键词向量 Wq，扩展后的关键词向量 We 为：

$$We = Wq \bullet P(W, W) \qquad \cdots\cdots \qquad (11)$$

Wq 是相当稀疏的，而 We 几乎在每一项上都有值。这是符合设计思想的，任何词之间（包含中英文词或其他语言的词）都有一定程度的语义联系，区别仅仅在于这种联系的强弱。

3 实验结果及分析

我们采用中英文双语文本对齐的语料库来进行实验，主要有以下两组数据：

a) 双语文本级对齐语料库 115 篇 新闻领域

b) 双语文本级对齐语料库 2041 篇 非特定领域

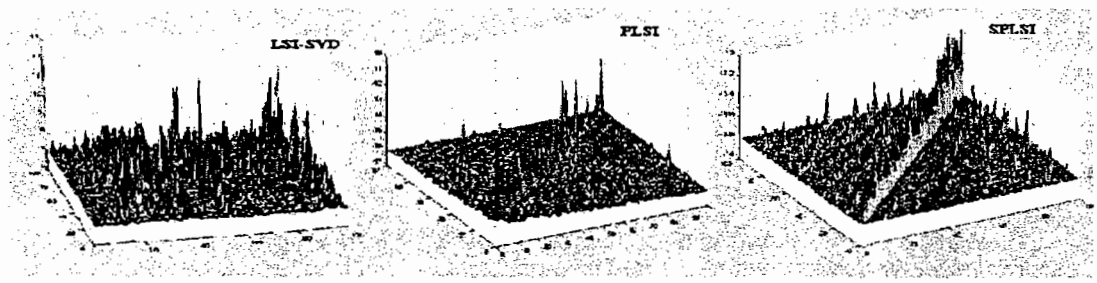
所使用的语料库包含在“中文语言资源联盟语料库”（www.chineseLDC.org）中。

3.1 多语言文本分类的对比实验结果

采用以上的两个语料库，分别构建 SPLSI 标引模型，然后应用所获得的词间相似度矩阵进行文本分类。同时，也实现了基于双语词典的关键词匹配（KW+Trans）等算法作为对比。

训练语料库	算法	主题维度	关键词数	集内测试集文档数	集内准确率	集外测试集文档数	集外准确率
双语文本对齐100篇新闻领域	KW+Trans	--	8043	25	100.0%	15	80.0%
	LSI-SVD	20	8043	25	92.0%	15	73.3%
	PLSI	20	8043	25	96.0%	15	80.0%
	SPLSI	20	8043	25	96.0%	15	80.0%
双语文本对齐1000篇非特定领域	KW+Trans	--	33045	38	89.5%	41	82.9%
	LSI-SVD	时间和空间复杂度太大，无法计算					
	PLSI	50	33045	38	92.1%	41	85.4%
	SPLSI	50	33045	38	94.7%	41	87.8%

对 SVD,PLSI,SPLSI 三种算法，用 100 篇双语文本分别训练出文档间相似度矩阵 *P(D,D)*：



图中，从左下到右上的对角线，显示了文档的自相似度值。显然，SPLSI 在这条线上的峰值更突出，体现了更强的物理意义；并且矩阵在整体分布上更为均匀，反映其分类更为合理。

3.2 跨语言文本检索的对比实验结果

采用双语对齐的 1000 篇文本作为训练集，分别获得 PLSI 和 SPLSI 词间相似度矩阵 *P(W,W)*。然后，随机选择若干组查询关键词，依据公式（10）在文本库中查找相关的文本。其中，KW+Trans 表示的是基于双语词典的关键词检索算法。SPLSI 算法的综合指标比其他算法更出色，尤其在召回率方面，充分体现了语义扩展的强大功能。

查询关键词	算法	集内准确率	集内召回率	集外准确率	集外召回率
中国贸易	KW+Trans	94.3%	61.1%	92.8%	56.5%
	PLSI	91.0%	86.2%	87.2%	83.1%
	SPLSI	92.7%	87.8%	90.3%	85.4%
football star match	KW+Trans	91.6%	63.7%	91.4%	58.3%
	PLSI	90.3%	88.2%	87.7%	84.7%
	SPLSI	92.5%	89.3%	91.9%	88.6%

在跨语言文本检索中，还有一个很重应用领域是：译文检索。采用与上述相同词间相似度矩

阵 P(W,W)，进行译文检索；集内测试集采用的是直译文本，集外测试集采用的是意译文本。

算法	集内测试集文档数	集内准确率（直译）	集外测试集文档数	集外准确率（意译）
KW-Trans	24 * 2	95.8%	18 * 2	83.3%
PLSI	24 * 2	91.7%	18 * 2	88.9%
SPLSI	24 * 2	95.8%	18 * 2	94.4%

由此可见，对于意译的文本，语义扩展算法的优势是相当明显的。换一个角度来看，词间相似度矩阵也可以作为翻译概率矩阵来使用。

3.3 跨语言检索中的关键词扩展

跨语言检索中的关键词扩展分为三个层次：翻译扩展、同义词扩展、语义扩展。前面反复用到的跨语言词间相似度矩阵，能够将这三个层次的扩展融为一体。利用公式（11），得：

查询词	扩展词
自然 语言	natural language NLP 语法 linguistic
computer hardware	计算机 电脑 硬件 system 显示器 配件 CPU

4 结束语

本文提出了一种隐含语义标引的新算法，称为“有指导的统计隐含语义标引”(SPLSI)。这种算法能够基于双语语料，构造出跨语言的词间相似度矩阵。和以前的算法相比，SPLSI有以下三点的改进：一是基于词义的聚类结果更具物理意义，效果更好；二是词义聚类的过程更容易控制；三是算法的时间和空间复杂度更低。

参考文献

[1] DEERWESTER S.,DUMAIS S.T.,FURNAS G.W.,LANDAUER T.K., and HARSHMAN R., Indexing by latent semantic analysis. Journal of the American Society for Information Science (1990).

[2] THOMAS HOFMANN, Probabilistic Latent Semantic Indexing ,Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval ,1999.

[3] Jianfeng GAO, Jian-yun NIE, Endong XUN, Jian ZHANG, Ming ZHOU and Changning HUANG, Improving Query Translation for Cross-Language Information Retrieval using Statistical Models, SIGIR, 2001.

[4] Jun ZHAO, The Approaches and a Framework of Translingual Text Processing, Chinese- Japanese Natural language Processing Proseminar (2nd), 2002.

[5] Lisa Ballesteros and W. Bruce Croft. Dictionary methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 791-801,1996.

[6] Lide Wu, Xuanjing Huang, etc., FDU at TREC-9: CLIR, QA and Filtering Tasks. In: The Ninth Text REtrieval Conference (TREC 9), 2000