基于流形排序的领域词抽取方法*

宋涛 李素建
北京大学计算语言学研究所 北京 100871
E-mail: songtao@pku.edu.cn lisujian@pku.edu.cn

摘 要：领域词通常是由一个或多个领域部件词组成的短语，其领域性主要由部件词体现。由此，本文收集领域文本，将其中候选短语构建成短语网，并提出假设：具有相同部件词的领域词之间具有紧密的联系，互相推荐。在此假设下，本文利用领域词的内在联系，引入基于流形的半指导排序方法，标记少量领域词，通过短语网将领域性分数进行传播，从而计算出所有短语的领域性分数，选取高分的短语作为领域词。我们在4个领域上进行了实验，结果表明该方法的有效性。

关键词：领域词，领域部件词，流形排序，半指导学习

Manifold-ranking Based Domain Phrase Extraction

Song Tao, Li Sujian
Institute of Computational Linguistics, Peking University, Beijing, China, 100871
E-mail: songtao@pku.edu.cn lisujian@pku.edu.cn

Abstract: Domain terms usually have one or more domain term components, and their domain importance is mainly up to their components. In our paper, several domain texts collected as corpus, we develop a weighted phrase network with candidate phrases as nodes in the corpus, based on the prior recommendation assumption, which means domain terms with the same component are likely to have similar importance scores. Under this assumption, we apply manifold-ranking semi-supervised learning on domain term extraction task. With only a few domain terms tagged, all phrases get their scores by spreading domain importance to their similar neighbors via the weighted network. The process is repeated until all scores are stable. Finally, we extract phrases with higher importance scores as domain terms. This method is experimented in four domains, and proves effective.

Keywords: domain term, domain term component, manifold-ranking, semi-supervised learning.

1 引言

上世纪九十年代以来，伴随着互联网的发展，各个领域的信息也呈现指数级增长的趋势。互联网为用户提供大量信息的同时，也给用户搜索有用信息带来困难。互联网上的信息主要是文本形式，数量过多过于通用且冗余严重，而用户需要的是领域性强、相对深入且精简的知识，可见领域性知识的研究具有重要的现实意义。领域性词汇是领域知识的基础，且对领域知识也具有一定的表示能力。因此，领域性词汇的研究是有必要的。中文处理方面，领域词汇的提取对汉语分词大有益处。对于词之间没有自然分割符号的汉语来说，分词往往是自然语言处理的第一步，分

* 本文承NSFC面上项目（项目号60875042和60773173）和NSFC重大研究计划/培育项目（项目号90920011）的资助

词词典则是实现准确分词的基础[1]。词典包含通用词典和领域词典，通用词典已相对完善，领域词典变化性大不易建立，传统人工编撰领域词典的方法需要领域专家投入大量的精力，且新词汇难以及时更新。领域词的自动提取可以弥补人工的缺点。领域词的自动提取还可以应用到很多领域，例如领域知识挖掘、领域本体构建[10]、信息检索、信息抽取、文档分类等，可见提高领域词提取的精准程度和自动提取效率将对自然语言处理有十分有益的帮助。

国内外已有一些关于自动获取领域词的研究。基于规则的方法提前获得或者建立规则模板，再与模板匹配。依靠领域专家手工收集规则模板费时费力，且随着规则集增大易出现不完备性和二义性，难以建立复杂全面的规则集合，现在使用较少。现在的研究主要是基于语料库和机器学习的方法。汉语方面有研究者提出一些在粗糙切分结果基础上提取领域词的方法。[1]提出基于切分单位的最大匹配算法，找到所有出现一次以上的短语，去除其中能被其他短语覆盖的短语，得到最长组合模式的短语。这种方法对语料依赖性大，不能识别单词型领域词。[2]中提出两个抽取原则NCI(normalized corpora impurity)和NCI(normalized corpora impurity)来衡量候选短语的领域性。[3,4]提出了术语部件的概念。对计算机领域术语部件的属性进行了描述，但主要工作还在于人工总结经验，移植到其他领域中还需要重新总结经验。[5]是一种不切分通过找领域词边界的方法，试图从切分结果词串中找到领域词之间的分隔符，从而找到领域词。但存在相邻分隔符之间通常不止包含一个领域词和分隔符的选取控制难等问题。已有研究对语料的利用通常只考虑词汇的统计量，对词汇之间的关系没有太多利用。本文试图通过基于流形排序的方法利用词汇之间的关系进行领域词提取。

基于图的方法能够利用节点之间的关联关系，广泛应用的如 PageRank 和 HITS。基于图排序的方法已经被成功的应用在多文档摘要上，[9]中利用 PageRank 算法思想，采用基于图排序的 LexRank 方法提取摘要。该算法将句子作为节点，节点间边的权重是句子节点的相似度，将句子之间的相似程度看作一种推荐。因此可以用迭代算法求得各个节点的分数，得分越高的句子重要性越大，抽取出来作为该主题的文档摘要。[11]中采用[7]中的基于流形的排序算法，处理以查询为焦点的多文档摘要提取问题。句子分数综合其偏向查询的程度和意义独立性，算法中通过相似度将查询的分数传播到句子中，选取得分高且信息量大的句子作为摘要。而短语与句子一样，也是若干词的有序集合，受图模型思想应用于摘要的启发，领域词的提取也可以采用类似的方法。

本文在切分基础上，展开对领域词自动提取的研究。领域词通常由通用词典中的一个或几个词组成，例如经济领域的“花旗银行”由“花旗”和“银行”构成，“花旗”和“银行”称为领域部件词。这里将领域词抽取分为领域词部件抽取和领域词抽取两部分工作。对于领域部件词，通过链接分析得到使用度以及不同领域的使用差异得到领域部件词[6]。本文的工作重点在领域词抽取上，提出一种基于流形排序[7]的领域词抽取方法。利用图排序算法思想，将切分后语料中 n 元短语作为候选短语，建立带权图，候选短语作为节点，句子节点的相似度作为边的权重，计算并选取领域度高的候选短语作为领域词。方法考虑了短语之间关系，利用半指导学习思想只需标记少量领域词，自动得到大量领域词。

本文其他部分组织如下：第二部分将界定本文工作的相关概念以及概括介绍整个方案。第三部分将详细介绍本文的方法。第四部分在多个领域的语料上进行实验，通过本文提出的领域词提取方法获得各个领域的领域词并评测实验结果。第五部分对全文进行总结，并展望进一步的研究工作。

2 涉及定义及方案设计

2.1 定义

定义 1：领域度指一个词或短语代表其所属领域的能力。

在本文中用词或短语的领域度得分表示这种能力的大小。满足如下假设，领域度高的词代表了相关领域的某个概念，领域词在其概念所代表的领域一般具有较高的分数。

定义 2：领域词指能够代表某个领域的某些概念的词或短语，具有较高的领域区分度，具有较高的领域度分数。

例如军事领域的领域词“国家导弹防御系统”由“国家”、“导弹”、“防御”和“系统”组成。需要说明的是，本文中领域词特指的是短语，领域部件词指的是词。

定义 3：领域部件词指具有领域性的单独的词，或者构成领域词的词。本文中规定领域部件词是切分单位。

例如，上例中军事领域的领域词“国家导弹防御系统”，由“国家”、“导弹”、“防御”和“系统”组成，这四个词都是领域部件词，其中“导弹”也是具有领域性的单独的词，本文对这两种领域部件词不作区分。

需要说明的是，领域部件词可能是通用词，例如上例中军事领域的领域部件词“系统”本身是通用词，并不具有很高的领域区分度，但是在军事领域中是领域词的成词部件，可以帮助构成一些军事领域词，例如“航空动力系统”等，在我们的研究方法中这种通用词对发现其他领域词时也有贡献作用，因此也作为领域部件词考虑。

2.2 方案设计

在对一个领域的语料建图的过程中，将候选短语看作节点，其中标记出少量领域词作为已标记节点，其他候选短语作为未标记节点，这样提取领域词任务可以看作一个半指导学习问题。用节点领域度表示短语的领域度，领域词提取就是选取领域度高的短语，则任务转化为一个节点领域度排序的问题。下面简述本文方法的流程。

本文的语料为领域文本，最终希望从中提取出领域词。主要流程分为三步：获取候选短语；应用基于流形排序的算法得到候选短语的领域度；排序并得到领域词。第一步在切分的基础上提取满足一定条件的短语作为候选短语，第二步标记少量领域词，并依此对所有候选短语赋初值，然后构建带权图，用相似矩阵表示边的权重。通过流形排序算法的迭代过程，节点将其领域度通过带权图传播到其他节点。迭代收敛后得到所有候选短语的得分。第三步排序并做后处理，最后得到领域词。

3 基于流形排序的领域词提取

3.1 候选短语选取

候选短语的选取是领域短语提取的基础。首先对语料进行切分和词性标注，本文实验采用的切分工具是中科院提供的开源切分工具。对语料中某个领域的文本进行切分标注后，得到包含

所有字符的以空格分隔的词和词性串 $w_1p_1w_2p_2...w_Np_N$（w_i 表示切分单位的词，包括标点；p_i 表示词 w_i 在句中的词性），从中提取 n 元词串作为候选短语。为了尽可能使提取的 n 元词串是有意义的短语，进行如下预处理：（1）以标点和部分停用词作为“词义分隔符”，将切分后的一个长的词串分隔成若干子段，这样做是因为本文假设“词义分隔符”左右两边的词在意义上是断开的，不能组成有意义的短语；（2）在所有子段中抽出所有 n 元词串，需要满足词串的最后一个词是动词或者名词，这样做也是为了增大 n 元词串是有意义短语的可能性。另外，在本文实验中 n 选取 2-4.，即候选短语的长度可能是 2、3 或者 4。经过上述处理，得到各个领域的候选短语。

3.2 流形排序算法

流形排序方法[7]是一种全局排序算法，来源于半指导学习方法[8]，最初用来对满足流形结构假设的数据进行排序。流形排序基于如下假设：（1）相近的节点倾向于有相近的排序分数；（2）相同结构（特别地被称为类或流形）中的节点倾向于有相近的排序分数。在领域词提取任务中，我们认为候选短语作为节点的图结构满足流形结构假设。直观上，相近的节点拥有一致的领域部件词，节点间相似度较大，在迭代过程中逐渐被赋予相近的分数，领域度分数接近则领域度排序相近，满足假设。

[7]中给出流形排序的直观描述。在数据集上构建带权图，对已标记点（本文方法中指已标记领域词）赋正值，对其他待排序节点赋零值，然后所有节点通过带权图将它们的分数传播给相邻近的节点。例如领域度高的节点可以将其领域度传播给与它含有相同词 t 的节点，并且若词 t 是领域部件词，则词 t 的传播作用更大。因此，可以说领域词节点将其领域度通过领域附件词传播到与其相似的节点，通过领域部件词加强其他节点分数，使得目标领域词节点分数逐渐增加并最终能够被抽取出来。迭代这种传播过程直到全局稳定状态，这样所有节点都得到最终的排序得分。本文中，候选短语作为节点，候选短语用词向量表示，每个维度是词的 *tf*(*term Frequency* 词频) • *ipf*(*invert phrase frequency* 倒排短语频率)值，节点相似度用向量余弦相似度计算。候选短语间相似度作为边的权重，下面给出本文采用的流形排序算法的形式化表示：

给定某个领域语料的所有候选短语 $\chi=\{x_1,...,x_m,x_{m+1},...,x_n\}$，其中 n 是候选词语总数目，m 是初始已标记的领域词，一般满足 $m<<n$，也就是说只需标记出少量的领域词。函数 $f:\chi\rightarrow R$ 定义一个排序函数，该函数对各个节点 x_i 赋予其排序分数值 f_i，表示该节点的领域度。f 可以看作向量 $f=[f_1,...,f_n]^T$。定义向量 $y=[y_1,...,y_n]^T$ 表示各个节点的初始值。

领域词提取算法如下：

算法	ManirankDomainPhraseExtraction(χ, y)
输入	候选短语集合 χ，候选短语初始值向量 y
输出	候选短语领域度分数 f^*
1	计算候选短语两两之间的向量余弦相似度 其中候选短语用词向量表示，每个词维度用 *tf* • *ipf* 值度量，候选短语 x 可以表示为向量 $x=(tf_{t1,x}\cdot ipf_{t1},...,tf_{tv,x}\cdot ipf_{tv})^T$，其中 V 是词总数。$tf_{t,x}$ 表示词 t 在候选短语 x 中出现的频率（次数），ipf_t 表示词 t 的倒排短语频率，这里用 $1+\log(N/n_t)$，其中 N 是短语总数，N_t 是有词 t 出现的短语的数目。给定候选短语 x_i 和 x_j，它们之间的余弦相似度用 $sim(x_i,x_j)$ 表示，用向量余弦公式计算如下：

$$sim(x_i, x_j) = \frac{\vec{x_i} \cdot \vec{x_j}}{|\vec{x_i}| \cdot |\vec{x_j}|} = \frac{\sum_{t \in (x_i \cap x_j)} \left(tf_{t,x_i} \cdot isf_t\right) \cdot \left(tf_{t,x_j} \cdot isf_t\right)}{\sqrt{\sum_{w \in x_i} \left(tf_{w,x_i} \cdot isf_w\right)^2} \cdot \sqrt{\sum_{v \in x_j} \left(tf_{v,x_j} \cdot isf_v\right)^2}}$$

2 构建候选短语相似度矩阵

将候选短语作为节点，候选短语之间的余弦相似度作为节点之间边的权值，节点之间的相似度大于 0，就在这两个节点之间产生一条以节点相似度作为权值的边。候选短语集合可以生成一个无向图 G 来反映短语之间的关系。定义对称矩阵 $S=(S_{ij})=(\mathrm{sim}(x_i,x_j))$作为候选短语相似度矩阵。需要注意的是，为了防止节点在第 4 步中的自加强，定义 $S_{ii}=0$.

3 相似度矩阵行归一化

定义对角矩阵 $D=(D_{ii})$,第 i 个元素表示相似矩阵第 i 行的和，即 $D_{ii}=\sum_j \mathrm{sim}(x_i,x_j)$. 对相似矩阵归一化为 $T = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$

4 迭代 $f(t+1)=\alpha Tf(t)+(1-\alpha)y$ 直到 f 收敛到 $f^*=(f_1, \ldots, f_n)$.

图 2::流形排序算法

算法中第三步对相似度矩阵行规一化保证了迭代算法的收敛性，[7]中证明节点分数最终能够收敛到 $f^*=(1-\alpha)(I-\alpha T)^{-1}y$，当 $\alpha \in [0,1)$，其中 I 是 n 维的单位矩阵。第二步中定义 $S_{ii}=0$ 保证各个节点在迭代过程中不会自加强。迭代过程中使用相似度参数 $\alpha \in (0,1)$ 来平衡邻近节点和初始排序分数对节点分数的影响程度，参数 α 值越接近 1 邻近节点对节点分数的影响越大，α 越接近 0 初始分数对节点分数的影响越大。

3.3 选取领域词

由于在得到候选短语时未作过多约束，得到的候选短语中有的会出现包含关系。例如经济领域的候选短语“非 寿险 业务”包含候选短语“非 寿险”等。在选取领域词时将被包含的短语去除后，依据领域度的排序结果选出领域词。

4 实验

4.1 实验数据和评测方法

为了减少不同领域偶然性的影响，实验选用 4 个不同领域的语料，分别为“军事”、“体育”、“经济”和“娱乐”，每个领域分别有 100 个文档，语料来源于新浪网站各个领域的网页。对结果的评测主要依据精确率和不同方法下结果的比较，实验中将各个领域排名最高的 200 个短语作为提取出的领域词，精确率是其中正确的领域词所占的比例。

需要说明的是，由于本文方法提取领域词时，主要考虑的是短语的领域度（termhood），没有对短语的成词性（unithood）进行判断和处理，这也是因为成词性是不易控制和检测的。因此，在本文中将包含领域词的短语也认为是正确的抽取结果。例如，抽取出的短语“不管 北朝鲜 危机”中包含领域词“北朝鲜 危机”，也认为这种提取是有意义的。这种情况的出现是由于没有将“不管”作为分隔词，在这里我们认为这并不影响它的正确性。因为我们抽取领域词的目的是让

读者能够更加高效的了解文本领域的大意，包含领域词的短语可以起到这样的作用。

实验主要为了完成两方面工作：（1）实验并测试相似度参数 α 对实验结果的影响。（2）与其他基于图模型的排序算法 PageRank 结果比较。证实本文的方法对领域短语的提取有积极的作用。

4.2 实验结果

在对相似度参数 α 的实验中，在一个领域的实验中，我们令 α 取 0.5-0.9 的以 0.1 为间隔的 5 个不同值，记录和观察多组实验的结果。下图是在经济领域中，α 取不同参数时的实验结果的精确率，其中横坐标表示 α 的不同取值，纵坐标表示在参数设定的情况下实验结果的领域词精确率。

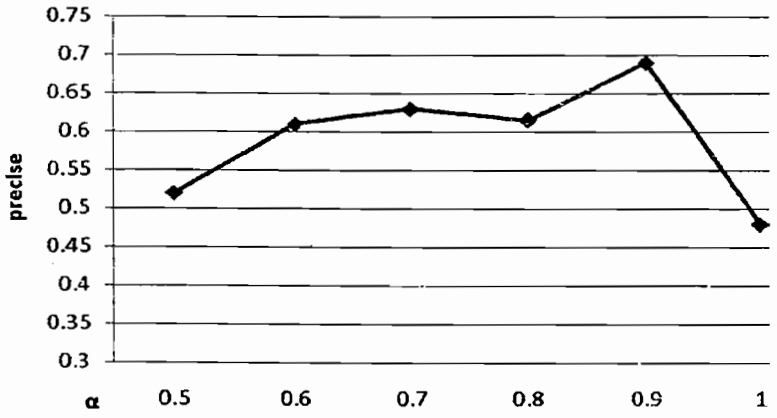

图 3:经济领域不同参数下的精确度

通过上表可以看出，在本文提取领域词的流形排序算法中，相似度参数对结果有影响。并且随着相似度参数 α 的增大，实验结果有提高。这说明领域部件词的传递作用起着积极的作用，对这种传递作用赋予的权重越大，结果越高。同时，也可以发现实验结果在 α 取 0.6 到 0.8 时变化幅度并不大。这可能是因为初始值影响与迭代作用影响的均衡所致。

当相似度参数 α 取 1 时，迭代公式的形式是 $f(t+1)= Tf(t)$，各个节点没有初始值，节点的分数由节点之间边的权重决定，迭代计算并在 f 稳定后停止迭代过程，与 PageRank 的迭代算法过程相同，因此相当于用 PageRank 算法计算节点的分数。因此，α 取 1 时的实验结果是 PageRank 算法下的实验结果。通过上图可以发现，本文中使用的基于流形排序的算法与 PageRank 算法相比，提取领域词结果精确率显著提高。可能的原因是在候选短语中，领域词所占的比例较低，使用 PageRank 算法时，对领域词和非领域词的贡献没有区分，非领域词对结果干扰较多。而本文所采用的算法有少量标记数据，对其中部分领域词予以较高的初始值，它们将分数通过领域部件词传播开来，使得实验受非领域词的干扰相对较少，因此结果更好一些。另外，结果还受语料质量和分词程序质量的影响。例如原文中“四大行人事变动”被切分成了“四大 行人 事变”等会带来不可预知的影响。

由于其他领域的结果与经济领域类似，本文中不再一一说明。只列举抽取结果中的个别领域词以供参考。例如经济领域的“银 监 局”、“外汇 市场”、“雷曼 兄弟”等，娱乐领域的“最

佳 女 主角 奖项”、“热门 金曲”、“北京 戏剧 学院”等，军事领域的“红 箭 表演队”、“多 功能 卫星”、“现代化 高 性能 潜艇”等，体育领域的“上海 大师赛”、“新 浪 体育”、“最佳 射手”等。

5 结论

本文将流形排序算法应用在领域词提取问题，算法将短语模型化为图，应用半指导学习的流形排序算法。方法的目的在于利用领域部件词的联系作用，通过少量的领域词发掘更多的领域词。实验证明该方法对领域词提取有帮助作用。在进一步的工作中，主要解决短语的成词性问题，克服初始标记领域词对结果的过分影响，参数的自动学习等。

参 考 文 献

[1] 孙霞，郑庆华，王朝静，张素娟. 一种基于生语料的领域词典生成方法. 小型微型计算机系统，2005，26(6):1088~1092.

[2] Tao Liu, Xiaolong Wang, Zhiming Xu, Qiang Wang. Domain-Specific Term Extraction and Its Application in Text Classification. Proceedings of 8th Joint Conference on Information Sciences , 2005, 1481~1484.

[3] 何燕，穗志方，段慧明，俞士汶. 一种结合术语部件库的术语提取方法. 计算机工程与应用，2006，42(33):4~7.

[4] 吴云芳. 信息科学与技术领域术语部件描述. 语言文字应用，2003, 34~39.

[5] Yuhang Yang, Qin Lu. Chinese Term Extraction Based on Delimiters. The Language Resources and Evaluation Conference Proceedings, 2008.

[6] 李素建等, 基于使用差异的领域性分析方法. 中文信息学报，2009, 73~78.

[7] Dengyong Zhou, Jason Weston, Arthur Gretton. Ranking on Data Manifolds. Neural Information Processing Systems, 2004.

[8] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, Bernhard Schölkopf. Learning with Local and Global Consistency. Neural Information Processing Systems(NIPS), Cambridge, MA, 2004. MIT press.

[9] Jahna Otterbacher, Günes Erkan, Dragomir R. Radev. Using Random Walks For Question-focused Sentence Retrieval, 2005. In proceedings of HLT/EMNLP 915~922.

[10] Yirong Chen, Qin Lu, Wenjie Li. Chinese Core Ontology Construction from a Bilingual Term Bank, The Language Resources and Evaluation Conference, 2008.

[11] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Manifold-ranking based topic-focused multi-document summarization. In Proceedings of IJCAI 2007.