基于短语网络的关键词自动提取方法 *

李广一，王厚峰

北京大学计算语言学教育部重点实验室，北京 100871

北京大学计算语言学研究所，北京 100871

E-mail: {liguangyi, wanghf}@pku.edu.cn

摘要：关键词自动提取是信息检索的一项重要任务。对于中文学术论文的关键词自动提取，本文提出了一种基于短语网络的排序方法。首先，利用 DF-AV 统计量提取关键词的候选短语，然后，以论文摘要为基础，构建短语网络，使用 TextRank 算法提取关键词，最后，利用基于 Perceptron 的重排序算法，进一步提升关键词提取的效果。本文在涵盖多个学科门类的论文集合上的实验表明，我们的方法是一种有效的关键词提取方法，，在测试集上前 5、前 10、前 15 个结果的 F 值分别为 27.96%、27.22%、24.07%。

关键词：关键词提取, DF-AV, 短语网络, TextRank, 重排序

Automatic Keyword Extraction Based on Phrase Network

Li Guangyi, Wang Houfeng

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, China

Institute of Computational Linguistics, Peking University, Beijing 100871

E-mail: {liguangyi, wanghf}@pku.edu.cn

Abstract: Keyword extraction is an important task for Information Retrieval. For the task of keyword extraction on Chinese theses, this paper presents a ranking method based on phrase network. First, extract keyword candidates by DF-AV. Second, build phrase network based on the abstract of the thesis and extract keywords by TextRank. Finally, improve keyword extraction with Perceptron reranking. Our experiments on various categories of theses prove our method effective. The top 5, top 10, top 15 F values on test data are 27.96%、27.22%、24.07% respectively.

Keywords: Keyword Extraction, DF-AV, Phrase Network, TextRank, Reranking.

1 引言

关键词是对一篇文档的内容和主题的浓缩，通常由几个词语或者短语构成。关键词在信息检索、文本分类、知识挖掘等领域有广泛的应用。在当今信息爆炸式增长的时代，关键词获取更是成为快速阅读海量数据的重要手段。

文本的关键词通常需要具有相关领域充足知识的人去获取，以保证关键词能够充分、准确地反映文档的内容和主题。然而,随着文档信息的大量涌现，已经无法单纯依靠人工进行关键词获取。因而，自动提取关键词已经成为自然语言处理和信息检索的一项重要研究任务。

目前，对关键词自动提取的研究主要面向新闻、学术论文、网页等文体。本文以中文学术论文作为研究对象，提出了一种基于短语网络的关键词提取方法，通过 DF-AV 统计量提取候选短语，基于短语网络使用 TextRank 算法提取关键词，并利用基于感知器的重排序，使关键词提取的准确性得到了进一步的提升。

2 相关工作

关键词通常为一个词或者多个词构成的短语，对关键词的提取可以分为两步：提取关键

* 本文受国家自然科学基金资助项目（No.61370117,61333018）、国家社科基金重大项目（No.12&ZD227）和国家 863 计划资助项目（No. 2012AA011101）资助

词候选集合和从候选集合中推荐关键词。

提取关键词候选的方法大致可以分为基于统计和基于规则的两类方法。基于统计的方法大多利用统计量来计算N元词串的紧密程度，以判断其是否是具有独立语义的短语，如均值和方差[1]、t测试[2]、卡方测试[3]、点对互信息[4]等。基于规则的方法主要利用了词语搭配的规则，如文献[5]提出英文关键词大多是名词性词组，“形容词+名词”是最常见的模式；文献[6]总结了汉语的短语结构规则。

推荐关键词的方法主要包括无监督的tf-idf[7]方法和基于有监督学习的分类或序列标注方法，如文献[8][9]等。近年来，一种基于图的排序算法Textrank[10]在关键词提取任务上取得了较好的效果，成为近期无监督关键词提取的主流方法，相关研究包括文献[11][12][13]等。此外，文献[14]采用了多种聚类方法提取关键词，文献[15]利用了隐含主题模型提取关键词。

3 学术论文关键词的特性分析

本文针对中文学术论文进行关键词提取的研究。为了自动提取关键词，我们首先对部分博士论文中的关键词及其特征作了分析。

本文使用了中国知网[1]上一定量的博士学位论文，包括论文的题目、摘要和关键词，领域涵盖了数学、物理、化工、天文、材料、矿业、法律、政治等多个学科门类，领域差异性明显。博士学位论文不仅学术水平高，更加符合写作规范，论文所给的关键词也相对严谨，可以用作关键词提取的参考答案。

我们随机抽取了1000篇博士学位论文，对其关键词的特性进行了分析。我们使用自行开发的基于Perceptron的分词和词性标注工具对文本进行预处理，该工具在人民日报1998年1月语料上分词的F值为96.30%，词性标注的准确率为95.10%。

3.1 关键词的出现频度

本文先对这1000篇博士论文的关键词的跨文档使用情况进行了统计。在1000篇博士论文中，共有5446个关键词，平均每篇论文有关键词5.446个。

我们对每个关键词在其对应的论文摘要中出现的频度进行了统计，统计的结果如表1所示：

频度	0	1	2	3-5	6-10	11-20	>20
百分比	16.62%	15.40%	10.67%	20.16%	14.14%	12.10%	10.91%

表 1 关键词论文内频度

可以看到，论文关键词的高频现象并不明显，在博士论文摘要内出现超过5次的关键词所占比例不足关键词总数的半数，有四分之一的关键词只出现了一次或两次，有六分之一的关键词甚至没有在论文摘要中出现。这说明，传统的基于词频的统计量方法可能并不适用于学术论文的关键词提取，因为其对于低频词的效果往往不理想。

3.2 关键词长度

为了发现关键词的长度规律，本文对所选论文的关键词的字数和词数分别进行了统计，统计的结果如表2和表3所示：

字数	1	2	3	4	5	6	>6
百分比	0.33%	14.01%	11.80%	41.70%	11.51%	10.28%	10.36%

表 2 关键词字数统计

[1] http://www.cnki.net

词数	1	2	3	4	>4
百分比	25.28%	51.58%	15.48%	5.03%	2.62%

表 3　关键词词数统计

从统计结果可以看出，关键词以4字2词为最多，单字的关键词极少，基本分布于长度2-6之间。从词的个数看，关键词中仅有四分之一为单个词语，四分之三是由多个词语组成的短语，因而，发现有独立意义的多词短语对关键词提取有重要的作用。

3.3　关键词词性

依据词性标注的结果，我们对关键词所包含的词语词性进行了统计，结果如表4所示：

词性	名词	动词	动词性名词	形容词	其他
百分比	53.42%	19.77%	6.11%	2.48%	18.22%

表 4　关键词词性统计

可以看出，关键词主要由名词和动词构成，但是，一些其他词性的词语也出现在个别关键词中，如标点和连词、助词等。我们还对多个词语构成的关键词的最后一个词的词性进行了统计，大多数情况下，末尾词是名词。但也只占58.96%，这低于我们的预期。其中，一个重要的原因是词性标注的错误，包括汉语中缺乏形态变化导致的错误，例如短语“关键词提取”，提取一词在这里是具有名词功能的动词，但其形态和作为动词时相同，而对于英语来说，“提取”一词作为动词时为extract，作为名词时为extraction，有显著的不同。因此，关键词的尾词不应局限于名词，英语的短语构成规则并不完全适用于汉语。

3.4　关键词词频及上下文词频

本文还对关键词中的词语词频和关键词在论文摘要中的上下文词频进行了统计，关键词中出现频率最高的五个词为法律、制度、社会、资源、司法，全部都是社会科学类的词汇。可以看出，社会科学类的论文关键词词频集中较为明显。而关键词在摘要中前后出现的词语中，标点符号、“的”、“对”、“了”、“和”这样的停用词出现的最多。通过进一步观察，我们发现，关键词中的词汇呈现显著的专业性特征，而关键词前后出现的词汇则具有一般性的特点。这对于关键词提取是一个重要的线索。

4　基于短语网络的关键词提取方法

在对学术论文关键词特性进行统计分析的基础上，本文提出了基于短语网络的关键词提取方法。基本流程是，首先，基于 DF-AV 统计量，从学术论文的题目和摘要中，提取关键词候选短语；然后，基于关键词候选短语，依据文档结构，构建由词汇和短语构成的词图，再利用 TextRank 对词汇和短语进行排序；最后，利用 Perceptron 算法，对 TextRank 排序的结果进行重排序，最后得到关键词提取的结果。

需要说明的是，由于本文采用的是抽取式关键词获取方法，而统计结果显示，有大约六分之一的关键词并未在题目和摘要中出现，这部分关键词是无法从文本中抽取到的。为了便于评测，本文只选取了关键词全部在题目和摘要中出现的博士学位论文进行实验。另外，由于所选论文的跨度大，关键词文档间共现较少，因而关键词提取均针对单一文档，并未考虑文档间的关系。

4.1　基于DF-AV的关键词候选短语提取

对关键词词长的统计结果显示，只有大约四分之一的关键词由单个词组成。如何将词语组合成短语，从中选出关键词的候选，就是关键词提取的一个很重要的问题。上文的统计分

析发现，有大量关键词并不符合名词性短语的构成规则（即短语的尾词为名词），因而，我们没有采用规则的方法，而采用了基于统计的方法。

候选短语选取的主要目标是选出联系紧密的词语构成短语，尽可能多地覆盖真实的关键词。因而，这一步的关键是保证召回率，在此前提下，尽可能减少非名词性短语及无实际意义的短语的干扰。

互信息常用于衡量词语间紧密程度的统计量，两个词语w_1,w_2间的互信息定义为$\mathrm{MI}(w_1, w_2) = \log \frac{p(w_1 w_2)}{p(w_1)p(w_2)}$。本文尝试了利用互信息提取候选短语：首先，计算相邻词语间的互信息，然后，对词语间互信息进行排序，选取阈值，最后，选出所有词语间互信息都高于阈值的短语，作为关键词短语候选。

但是，我们发现，通过互信息获取短语候选的效果并不理想，召回率低且包含了大量无意义短语。其中的一个重要原因在于，论文摘要的文本短，词频低，仅仅基于频率的互信息难以准确反映词语间的紧密程度。

AV统计量[16]是对于中文新词提取有良好效果的一种统计量，由于关键词候选短语的提取和中文新词提取具有类比关系，本文利用AV统计量进行关键词候选短语提取。我们定义S_L为短语phr左侧出现的词语集合，S_R为短语phr右侧出现的词语集合，于是，短语phr的左AV统计量$\mathrm{AV}_{\mathrm{phr}} = \mathrm{sizeof}(S_L)$，短语phr的右AV统计量$\mathrm{AV}_{\mathrm{phr}} = \mathrm{sizeof}(S_R)$。AV统计量越大，说明该短语在构成上独立性越强，就越有可能是候选短语。针对短语phr，我们引入了一个紧密度度量公式$\mathrm{Score}_{\mathrm{phr}} = \mathrm{Freq}(\mathrm{phr}) \times \mathrm{AV}_L(\mathrm{phr}) \times \mathrm{AV}_R(\mathrm{phr})$。任何只出现1次的短语得分都是1，而其中大多数为无意义短语，因此，选取所有分数大于1的短语作为候选短语。

利用AV统计量的提取方法能有效地提升关键词的召回率，但是对于只出现一次的关键词，仍然无法提取。为了进一步提高召回率，本文结合各个可能候选所在上下文词的情况进行了扩展。上文的分析表明，大量关键词的上下文词语都是常用词，可以利用这一特点扩展关键词候选。通常认为，常用词可以利用文档频率(DF)来反映，越常用的词语，越可能在更多的文档中出现，因而其文档频率就会越高；反过来，关键词的文档频度就相对较低，如果一个词或短语的前后词语文档频率较高而其自身的文档频率不高，则很可能就是一个关键词候选。基于该假设，本文提出了利用DF-AV统计量提取关键词候选短语。

不同于AV统计量计算短语前后不同词语的个数，DF-AV计算短语前后不同词语的DF之和。基于中文gigaword语料，我们统计了词语的文档频数。为了计算方便，我们使用了文档频数而并没有使用频率，由于gigaword语料规模很大，我们对文档频数进行了取对数处理。于是，定义DF-AV和短语的分数如下：

$$\mathrm{DFAV}_L = \sum_{d \in S_L} \log \mathrm{DF}(\mathrm{phr})$$

$$\mathrm{DFAV}_R = \sum_{d \in S_R} \log \mathrm{DF}(\mathrm{phr})$$

$$\mathrm{Score}_{\mathrm{phr}} = \mathrm{DFAV}_L(\mathrm{phr}) \times \mathrm{DFAV}_R(\mathrm{phr})$$

由于超过4个词的短语作为关键词的情况非常少，因此我们限定短语的最大词数为4。对于所有词数小于等于4的短语，计算分数，选取分数大于某一阈值的短语作为关键词候选。阈值的选择可以实现召回率和短语候选总数之间的平衡。为了进一步减少无意义短语的干扰，选择候选短语时不考虑包含了标点、介词、助词、连词、代词的短语。实验结果显示，利用DF-AV统计量提取的关键词短语候选效果最好，本文将在该结果基础上，进行后续的处理。

4.2 基于短语网络的TextRank

TextRank是2004年Mihalcea和Tarau受PageRank算法[17]启发提出的基于图的排序算法。

TextRank将文档看作一个词的网络，网络中的链接表示词和词之间的关系，一个词的重要程度由链向它的其他词的重要性决定。定义有向图G=(V,E)为该词的网络，V为节点集合，E为有向边的集合，对节点V_i，定义$In(V_i)$为有边指向V_i的节点集合，$Out(V_i)$为V_i指向的节点集合，则节点V_i的得分定义为：

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_i)|} S(V_j)$$

TextRank可以通过迭代或矩阵运算得到稳定状态下每个节点的得分，依据得分的排序，可以得到关键词提取的结果。在以往的TextRank用于关键词提取的研究中，通常构建词汇的网络，然后依据得到的结果来构建短语，这种做法无法保证生成有意义的短语，并且真正的关键词中也无法保证所有词语都具有较高的重要性。因此，本文提出构建基于短语的网络，以此为基础，通过TextRank算法，直接产生关键词的排序结果。

本文利用DF-AV统计量提取了关键词短语候选，基于这些短语候选，可以构造短语的网络。TextRank一般以词语在窗口内的共现关系作为两个词语间存在链接的依据，本文将这种关系扩展到短语之上。以“辐射带 电子 通量 模式 研究”为例，假定候选短语中包含了“电子通量”、“通量模式”、“电子通量模式”，则将这些短语和词汇一起构造词汇连通图，如果两个词汇或者短语相邻，那么就在二者之间连接一条边。据此构建的词汇连通图如图1所示：

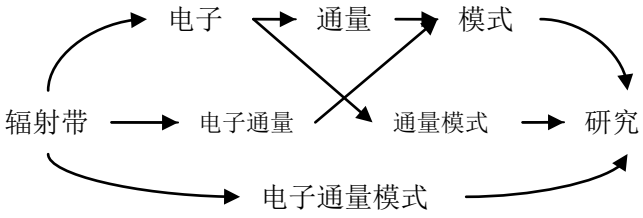

图 1　词汇连通图

然后，基于词汇连通图，根据所选窗口大小，建立TextRank的词汇网络。具体方法是，假如窗口大小为n，如果两个节点之间存在一条长度小于n的有向路径，那么就在两个节点间添加一个链接，重叠的两个节点间不会有链接，比如“电子”、“电子通量”、“电子通量模式”之间都不会有链接。以窗口大小取2为例，针对图1的词汇连通图构建的TextRank词汇网络图如图2所示：

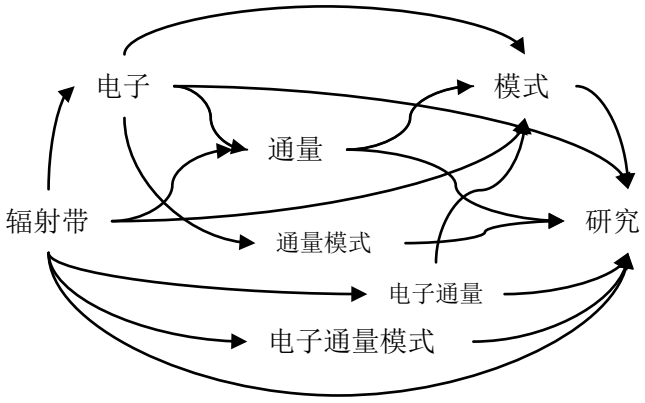

图 2　TextRank词汇网络图

在添加了短语后，词汇网络图的结构变得非常复杂。为了控制词汇网络图的复杂程度，

我们规定，如果候选短语phr中包含了词语w，且count(w)=count(phr)，即w和phr出现的次数相同，那么phr只作为w的部分出现。我们认为词语w在文档中并不单独表达语义，在构建词汇连通图时，涉及到候选短语phr的部分不再将词语w作为一个节点。例如，假定count(通量)=count(电子通量)，那么在构建词汇连通图时，包含“通量”这一节点的路径将被删除，从而使TextRank词汇网络图得到简化。

为了表征不同词汇的重要程度不同，本文采用了带权重的TextRank算法，为边e_{ij}设定权重w_{ij}，则节点V_i的得分定义为：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_i)} w_{jk}} WS(V_j)$$

w_{ij}主要考虑以下几个因素：

(1) 词语间距离。定义$dist_{ij}$为两节点在词汇连通图中的最短路径长度。定义$w_{dist} = 0.5 + \frac{0.5}{dist_{ij}}$。距离越长权重越小。

(2) 短语长度。依据关键词字数统计结果，依据Vj的长度设定$w_{len} = 0.5 + \frac{0.5 \times n_{len(V_j)}}{n_4}$，其中$n_k$表示长度为k的关键词频数。

(3) 短语候选得分。将短语候选的得分作为权重，短语得分越高则权重越高。定义$w_{score} = \sqrt[3]{score_{V_j}}$。

w_{ij}为上述权重的乘积，即$w_{ij} = w_{dist} \times w_{len} \times w_{score}$。

我们使用迭代法计算每个节点的重要程度，待算法收敛后，得分最高的节点对应的短语，即为通过TextRank算法提取的关键词。

4.3 基于Perceptron的重排序

对关键词的统计结果显示，关键词的词语构成和关键词在文档中上下文的词语构成具有一定的规律性。受此启发，我们在TextRank得到的关键词提取结果基础上，采用有监督学习方法，对提取结果进行重排序，以进一步提升关键词提取的效果。

为了适应关键词提取任务，我们对文献[18]提出的Generalized Perceptron进行了扩展，传统的Perceptron算法只求最优解，我们将其扩展为取top k个最优解。扩展后的Perceptron算法描述如下，定义TextRank提取出的关键词集合为S_T，对关键词$x \in S_T$，定义x对应的特征向量为$\Phi(x)$，参数向量为$\vec{w}$，则x的得分为

$$Score(x) = \Phi(x) \cdot \vec{w}$$

定义该文档的关键词个数为k，对于所有$x \in S_T$，Score(x)中第k大的为$score_k$，则算法输出的关键词集合S_{output}为

$$S_{output} = \{x | Score(x) \geq score_k, x \in S_T\}$$

定义该文档的真实关键词集合为S_{gold}，则Perceptron的参数更新策略为

$$\overrightarrow{w_{t+1}} = \overrightarrow{w_t} + \sum_{x \in S_{gold}} \Phi(x) - \sum_{x \in S_{output}} \Phi(x)$$

基于对关键词的统计分析，我们选取了以下的特征模板：

(1) 关键词短语

(2) 关键词包含的词语集合

(3) 关键词的首末词语
(4) 关键词的文档中出现位置前的词语集合
(5) 关键词的文档中出现位置后的词语集合
(6) 离散化的DF-AV短语得分
(7) 离散化的TextRank关键词重要性得分

5 实验及结果分析

5.1 实验

我们从多个学科门类的论文摘要中随机抽取500篇作为测试语料，将论文的题目和摘要作为文档，以论文原作者给出的关键词作为参考答案。我们的方法不考虑在摘要中未出现的关键词，我们在抽取实验数据时，只选择了含所有关键词的论文。

关键词提取结果的评价通常计算前5个、前10个、前15个结果的准确率、召回率和F值作为评价指标。计算多篇文档的指标平均值，有宏平均(macro-average)和微平均(micro-average)两种方式。本文在对实验结果评价时，同时计算了宏平均和微平均，以二者的平均值作为最终结果的准确率、召回率和F值。

关键词候选短语提取的目标是尽可能多地涵盖关键词，本文以文档平均召回率作为评价指标做了比较。我们采用了互信息、AV统计量和DF-AV统计量三种方法进行关键词候选短语提取，结果如表5所示：

方法	互信息	AV	DF-AV
平均召回率	55.04%	75.49%	84.10%

表 5 关键词短语候选提取结果

可以看出，本文提出的DF-AV统计量用于提取关键词短语候选提取效果最好。事实上，如果降低选取的阈值，DF-AV方法可以达到更高的召回率。但是随之引入的无意义短语也会增多，需要选择适当的阈值在二者之间实现平衡。我们选择的阈值为100。

在DF-AV提取的候选短语基础上，再构建词汇网络，并通过TextRank算法进行关键词提取。本文采用不同的权重设定进行关键词提取实验，实验结果如表6所示：

权重设定	F值 (Top 5)	F值 (Top 10)	F值 (Top 15)
设定为1	26.21%	25.79%	22.84%
w_{dist}	26.33%	25.78%	22.75%
$w_{dist} \times w_{len}$	26.37%	24.79%	22.21%
$w_{dist} \times w_{len} \times w_{score}$	27.34%	26.66%	23.76%

表 6 TextRank实验结果

表7显示，我们选择的边权重方法对TextRank的关键词提取效果有大约1%的提升，但是由于图的复杂程度较高，修改权重对提取效果的提升幅度比较有限，提取效果更多地取决于图的结构。

前两步都是无监督学习范畴，而基于Perceptron的重排序是有监督学习，因而我们另行抽取1000篇论文作为训练语料。我们对训练语料使用前两步的最好方法进行关键词提取，以提取的结果为基础，利用我们给出的特征模板，使用Perceptron算法进行训练。实验中，我

们选定迭代次数为30次。利用训练的模型，对测试语料TextRank的结果进行重排序，按照Perceptron算法的得分由高到低，重新输出关键词的排序结果。重排序的实验结果如表7所示：

	Top 5			Top 10			Top 15		
	P	R	F	P	R	F	P	R	F
重排序前	28.60%	26.64%	27.34%	20.71%	38.25%	26.66%	16.31%	45.01%	23.77%
重排序后	29.24%	27.25%	**27.96%**	21.13%	39.07%	**27.22%**	16.50%	45.63%	**24.07%**

表 7　Perceptron重排序结果

可以看出，基于Perceptron的重排序算法在各项指标都一定的提升，尤其是召回率的提升较为显著，说明重排序使一部分真实的关键词的排序得到了提升。这说明我们提出的基于Perceptron的重排序算法是一种有效的改进关键词提取效果的方法。但我们对结果的观察发现，重排序对于一小部分关键词仍存在负作用，重排序的稳定性还有待于提升。

我们的算法最好结果与传统tf-idf方法的比较数据如表8所示：

	Top 5			Top 10			Top 15		
	P	R	F	P	R	F	P	R	F
tf-idf	16.72%	15.67%	16.03%	14.28%	26.64%	18.44%	12.69%	35.39%	18.55%
我们的方法	29.24%	27.25%	**27.96%**	21.13%	39.07%	**27.22%**	16.50%	45.63%	**24.07%**

表 8 方法对比

可以看出，我们的方法在大多数指标都超越tf-idf方法10%以上，这说明，我们提出的方法是一种有效的关键词提取方法。

5.2　实验结果分析

上述实验数据说明，本文所提的方法对中文学术论文的关键词提取任务，取得了良好的效果。然而，我们的实验结果仍然偏低，主要有三方面的原因：第一，一篇论文平均关键词数目为5.4个，这就决定了top 10、top 15的准确率上界非常受限；第二，我们直接采用了论文原作者给出的关键词作为“标准答案”，关键词本身受到人为因素的影响，对同一篇文章，不同的人所给的关键词在词形上可能就存在很大的差异，而我们在评价时只有完全相同才认为提取的关键词正确；第三，由于跨领域的分词和词性标注还不够成熟，在学术论文上的分词和词性标注错误仍然普遍存在，对实验效果也会产生影响。

本文提出的方法仍有很多需要完善之处，比如关键词候选短语提取遗漏了部分关键词；构建TextRank的短语网络时仅考虑了窗口内共现信息，没有考虑依存关系或者语义信息；尽管我们选取语料包含了多个学科门类，但能否对各种学科的论文都具有良好的实验效果还有待于进一步验证。而且，我们忽略了未在文档中出现的关键词，如何实现对这类关键词的提取，也是值得研究的问题。另外，利用论文间的关联和领域知识，也是我们未来研究的方向之一。

6　结语

本文提出了一种基于短语网络的论文关键词自动提取算法，首先，利用我们提出的DF-AV 统计量提取关键词候选短语；然后，构建短语网络，设定边权，利用 TextRank 算法，

自动提取关键词；最后，我们利用基于 Perceptron 的重排序算法，对关键词提取结果进行优化，进一步提升了关键词提取的效果。实验结果表明，我们的方法是是一种有效的提取论文关键词的方法。

下一步，我们将继续深入研究，优化关键词候选短语提取方法，优化词汇网络的构建，并探究如何利用论文间的关联和领域知识，进一步提升关键词自动提取的效果。

参考文献

[1] Smadja F. Retrieving collocations from text: Xtract. Computational Linguistics, 1993, 19(1):143–177.

[2] Church K W, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990, 16(1):22–29.

[3] Church K, Gale W, Hanks P, et al. Using Statistics in Lexical Analysis. Lexical acquisition: exploiting on-line resources to build a lexicon, 1991. 115–164.

[4] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993, 19(1):61–74.

[5] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Proceedings of EMNLP, 2003. 216–223.

[6] 詹卫东.面向中文信息处理的现代汉语短语结构规则研究.北京大学,1999.

[7] Manning C, Raghavan P, Schtze H. Introduction to information retrieval. New York, NY, USA: Cambridge University Press, 2008.

[8] Turney P D. Learning Algorithms for Keyphrase Extraction. Information Retrieval, 2000, 2(4):303–336.

[9] Zhang C. Automatic keyword extraction from documents using conditional random fields[J]. Journal of Computational Information Systems, 2008, 4(3): 1169-1180.

[10] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts. Proceedings of EMNLP, 2004. 404–411.

[11] Wan X, Xiao J. CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction. Proceedings of COLING, 2008. 969–976.

[12] Zha H. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. Proceedings of SIGIR, 2002. 113–120.

[13] Ercan G, Cicekli I. Using lexical chains for keyword extraction. Information Processing Management, 2007, 43(6):1705–1714.

[14] Liu Z, Li P, Zheng Y, et al. Clustering to find exemplar terms for keyphrase extraction. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 257-266.

[15] Liu Z, Huang W, Zheng Y, et al. Automatic keyphrase extraction via topic decomposition. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 366-376.

[16] Feng H, Chen K, Deng X, et al. Accessor variety criteria for Chinese word extraction. Computational Linguistics, 2004, 30(1): 75-93.

[17] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: Bringing order to the web. Technical report of Stanford Digital Library Technologies Project, 1998.

[18] Collins M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 1-8.