情 报 学 报 ISSN 1000 - 0135
第 32 卷 第 8 期 797 - 804, 2013 年 8 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 32 No. 8, 797 - 804 August 2013

doi:10.3772/j.issn.1000-0135.2013.08.002

一种基于加权网络和句子窗口方案的信息检索模型[1)]

陆 伟 程齐凯

(武汉大学信息资源研究中心,信息检索与知识挖掘研究所,武汉 430072)

摘要 经典的信息检索模型在文档表示上多采用词袋模型,与此不同,本文提出了一种基于加权网络的信息检索模型。在这一模型中,文档被表示为一个加权共词网络,词汇在文档中的重要性通过词项节点在网络中的重要性加以衡量。基于固定窗口平移和句子窗口方案,本文提出了文本游走模型 Textrank 的四个检索模型变种,分别是 Win_Weighted_Textrank、Sent_Weighted_Textrank、Win_Weighted_Posrank 和 Sent_Weighted_Posrank。在 Reuter RCV1 上的实验证明,与无权网络模型 Textrank、Posrank 相比,本文提出的模型能显著地提升检索效果。

关键词 信息检索 信息检索模型 加权网络 共词网络

An Information Retrieval Model Based on Weighted Graph and Sentence

Lu Wei and Cheng Qikai

(*Center for Studies of Information Resources*, *Wuhan University*, *Wuhan* 430072)

Abstract A standard approach to Information Retrieval (IR) is to model text as a bag of words. Alternatively, this paper proposes a weighted graph based information retrieval model which expresses document as a weighted co-words network. We measure the "eliteness" of a term by weighted textrank. With the setting listed above, this paper's work contains two parts: we use weighted co-words network to represent a document instead of unweighted network; we use both fixed and dynamic co-occurrence windows to build co-words network. In this paper, we proposed four variants of Textrank, which are Win_Weighted_Textrank, Sent_Weighted_Textrank, Win_Weighted_Posrank and Sent_Weighted_Posrank. Experimental results on Reuter RCV1 show that methods we propose in this paper perform better than Textrank and Posrank, which are unweighted network models.

Keywords information retrieval, information retrieval model, weighted network, co-words network

1 引 言

信息检索模型是信息检索研究的核心内容之一。经典的检索模型包括向量空间模型[1]、概率模型[2]、推理网络模型[3]、语言模型[4,5]等。经典模型有着各自不同的数学基础,但却得到了非常类似的排名计算公式。之所以如此,原因之一在于这些检索模型多采用词袋模型作为文档表示方法。随着复杂网络,特别是语义网络的兴起,研究者开始探索文档的词项网络表示方法。既然文档可以通过词项网络表示,那是否可以将这一文档表示方法应用于信

收稿日期:2012 年 9 月 3 日

作者简介:**陆伟**,男,1974 年生,武汉大学信息管理学院,博士生导师,教授,主要研究方向:信息检索、知识挖掘、数字图书馆等。E-mail:reedwhu@gmail.com。**程齐凯**,男,1989 年生,武汉大学信息管理学院情报学系,博士研究生,主要研究方向:信息检索、数据挖掘。E-mail:chengqikai0806@gmail.com。

1) 本文系教育部人文社会科学基地重大项目"面向细粒度的网络信息检索模型及框架构建研究"(项目编号:10JJD630014);国家自然科学基金面上项目"基于语言模型的通用实体检索建模及框架实现研究"(项目编号:71173164)的研究成果之一。

息检索呢?

网络的概念在信息检索研究中很早就得到了关注,最早的研究可以追溯到 Minsky[6]。Turtle、Croft、Rijsbergen 等发展的推理网络模型也可以被认为是网络在信息检索中的应用[3,7]。近年来,研究者利用推理网络模型的框架,发展出一些基于文档或文档片段网络的检索模型:Kazai 等提出一种基于文档网络图的检索模型,这一模型将文档表示成树形结构,每个节点表示文档的片段,节点之间的边表示文档部件的关系,文档节点的权重根据节点的语言性质进行度量;Ogilvie 等提出了类似的模型,但在节点权值的设定上采用了类似于语言模型的方法;Quintana 等提出了 HRS 方法,将文档和查询都表示成概念网络图,将文档和查询的相关性度量问题转化为图匹配问题;Thammasut 等将随机游走应用于文档网络和词汇网络,文档与查询的相关度通过文档节点的可达性确定[8]。

将文档表示为词汇网络,首先由 Mihalcea 在 Textrank 模型中提出[9]。Blanco[10,11]将 Textrank 应用于信息检索,提出了一种基于文档词汇网络的信息检索模型。该模型设定在一定窗口内共现的词汇具有关联性,然后基于词共现关系构建文档的无权共词网络。在这一文档表示方法下,词汇在文档中的重要性可以通过词汇节点在共词网络中的重要性加以衡量,可用的词汇节点重要性衡量指标包括词项节点的 PR 值(pagerank)、节点的度(degree)等等[12]。然而,该模型将文档表示成无权网络,没有考虑边权对检索效果的影响;同时,这一模型采用窗口平移的方法构建词项网络,也没有很好地反映文本的语义特征。

本文在 Blanco[10,11]工作的基础上,提出了一种基于加权网络的信息检索模型,其基本思想是将文档表示为加权词项网络,将词项在文档中的重要性度量转化为词项节点在词项网络中的重要性度量:词项节点在网络中的重要性程度越高,对应的词项在文档中重要性越高。本文后续内容结构如下:第二节介绍模型的具体实现,第三节给出实验结果,最后对文章做总结。

2 研究方法

信息检索模型至少包含两项内容:文档和查询的表示以及相关性计算方案。前者通过某种转化方法,将文本和查询表示为特定的形式,以供相关性计算使用;相关性计算方案是信息检索模型的核心内容,它针对查询和文档的特定表示计算查询和文档的相关性得分,并返回一个按照分值排好序的文档列表。本节接下来在第一部分将介绍加权共词网络的构建方法,然后介绍了词汇重要性的度量方案,最后给出了查询相关性计算方法。

2.1 加权共词网络的构建

Blanco 等[11]使用词性无关词项网络和词性相关词项网络两种形式表示文档。共词网络基于窗口平移方法构建,即认为在给定窗口长度内的词汇具有共现关系。Blanco 等[11]构建的网络是无权网络,窗口大小也需要人工设定。本文将无权网络扩展为加权网络;同时,在共现关系的确定上,同时使用句子和固定长度作为窗口。

2.1.1 词性无关加权共词网络

词性无关共词网络利用词项的共现关系构建共词网络,每个节点表示一个词项,节点间边权表示两个词的共现频次。构建共词网络时,首先为文档中的每一个词构建节点;然后为节点构造边。本文使用了两种边构造方法,一种利用到基于固定长度窗口的词共现关系[9],另一种则利用词在句子中的共现关系构建网络。

第一种方法利用固定大小窗口的平移得到共词网络:给定长度为 N 的窗口,从文档词序列的首位开始向后平移,每次移动一个词汇的长度,在窗口内共现的词汇之间即需要构造边。具体的,给定窗口内共现词汇 A 和 B,且 A、B 为不同的词汇,A、B 对应的词汇节点为 NodeA、NodeB,Link(A,B)表示 NodeA 和 NodeB 之间的边,W(A,B)表示 Link(A,B)的边权,如果 Link(A,B)不存在,则为 NodeA,NodeB 构造 Link(A,B),令 W(A,B)=1,如果 Link(A,B)存在,令 W(A,B)= W(A,B)+1。

第二种方法同第一种方法类似,将句子作为窗口单位,每次平移的距离为句子长度,也就是说,对于文档 $D_i=\{s_0,s_1,\cdots,s_n\}$,如果词汇 A 和 B 在句子 s_i 中共现,且 A! =B,则为词汇 A 和 B 对应的节点 NodeA 和 NodeB 构造边,构造方法同前一种。

图 1 给出了一个基于句子窗口的词性无关网络的图形样例,根据路透数据集 RCV1 中编号为 250662 的文档构建网络并绘制。在构建网络前,对文档文本做了一些预处理工作,包括词干提取、停用词处理。250662 号文档在大多数检索模型中都被

判定为查询“Economic espionage”的最相关文档，在图 1 中可以看到，查询式包含的两个词汇都在共词网络中占据着较为中心的位置。

基于窗口平移和基于句子窗口两种方法各有利弊。根据乔姆斯基的句法理论，句子在语义表述上构成了一个基本的单位，句子内的词汇即使乱序，也能在一定程度上体现出句子的语义含义[13]。利用句子作为窗口构造共词网络可以较好的反映出原始的语义信息。但是，以句子为窗口构建共词网络某些时候也会出现网络不连通的问题，为后续的网络拓扑性质计算带来困难。因此，在文档长度较小的情况下，使用基于窗口平移的网络构建方法是比较合适的。

2.1.2 词性相关加权共词网络

词性相关共词网络的构建方法同词性无关共词网络的构建方法类似，也有着基于窗口平移的和基于句子的两种不同模式。两者区别在于：对于共现词汇 A 和 B，词性无关共词网络构建一个无向边 Link(A,B)，而词性相关共词网络则需要根据 A 和 B 的词性构造有向边。

计算语言学理论认为，词性不同，词汇在句子中的地位也会有所不同。Jespersen[14] 将词分成三种级别，级别的高低意味着词汇之间存在着附属或者修饰的关系。遵从 Blanco 等[10] 的原始设定，本文根据词性的不同将词分成四个级别，名词属于第一级别，形容词和动词属于第二级别，副词属于第三级别，其他词性的词汇属于第四级别。例如：

例 1：Some(4) pretty(3) good(2) cars(1)

在例 1 中，cars 是名词，属于第一级别，good 是形容词，修饰 cars，属于第二级别，pretty 修饰 good，属于第三级别，Some 属于第四级别。

词性相关网络的构建规则是：高级别的词汇对应的词项节点指向低级别的词汇对应的词项节点，相同词性的词汇对应的节点之间相互指向。给定共现词汇 A 和 B 及对应的词项节点 NodeA 和 NodeB，$C=\{1,2,3,4\}$ 表示词的词性级别，$Pos(C,A)$ 表示词 A 到词性集合 C 的映射。如果 $Pos(C,A)>Pos(C,B)$，则对节点对 < NodeA, NodeB > 按照如下规则处理：①如果 NodeA 和 NodeB 之间不存在边，则构造从 NodeA 指向 NodeB 的有向边 Link(A,B)，边权 W(A,B) 设为 1；②如果 Link(A,B) 存在，则令 W(A,B) = W(A,B) + 1；如果 $Pos(C,A)=Pos(C,B)$，则为节点对 < NodeA, NodeB > 和 < NodeB, NodeA > 分别根据规则①、②进行处理。

图 2 给出了一个词性相关共词网络的图形实

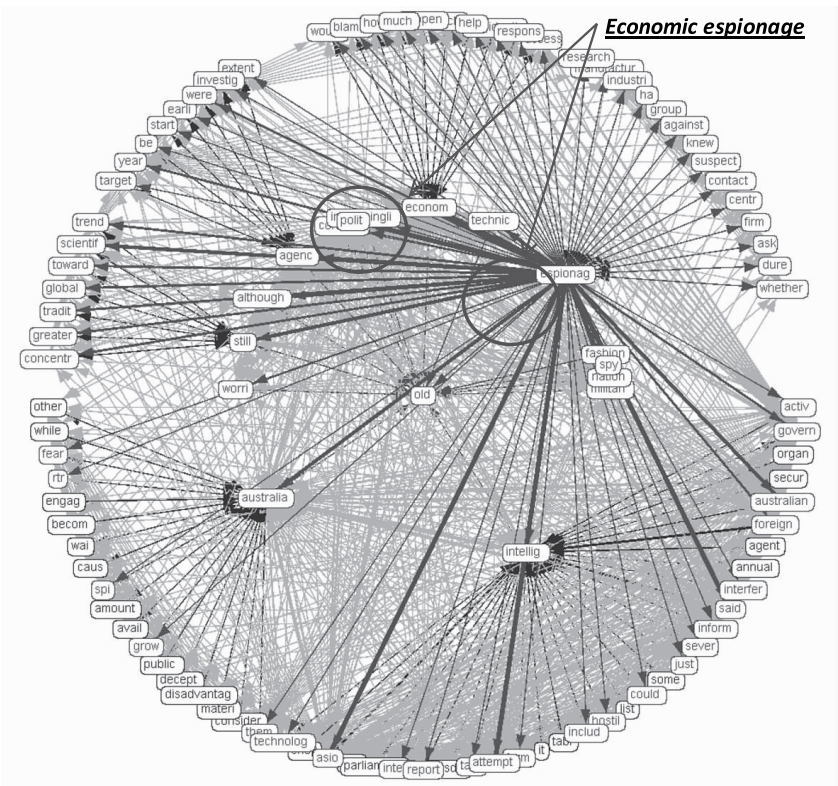

图 1　基于句子窗口的词性无关网络的图形(依据 RCV1_250662 号文档绘制)

例。同图 1 一样,图 2 基于 RCV1_250662 号文档构建,文档文本也使用了同样的预处理方法。同图 1 相比,图 2 的社区特征较为明显,相同词性的词汇聚集在一起,低级别的词汇较高级别的词汇有着更高的中心性。

2.2　节点重要性的度量

Pagerank 是链接分析的基本方法,起源于 web 分析领域[9, 12]。PR 值的计算公式见公式(1)。

$$S(v_i) = (1 + \phi) + \phi \sum_{j \in V(v_i)} S(v_j) / | V(v_j) | \quad (1)$$

$S(v_i)$和 $S(v_j)$表示节点 v_i和节点 v_j的 PR 值,$V(v_i)$和 $V(v_j)$分别表示指向 v_i 和 v_j 的节点集合,概率 ϕ 被称为阻尼系数,用于调节随机跳转概率。应用于词项网络,公式(1)构成了 Mihalcea [9]提出的文本游走模型 Textrank 的一个基础实现。

在共词网络中,词汇节点的关系强度不是均匀的,无权网络抛弃了网络的部分细节信息。因此,需要考虑加权网络的随机游走模型实现,本文使用公式(2)计算加权网络的节点权值 wpr。

$$S(v_i) = (1 - \phi) + \phi \sum_{j \in V(v_j)} \frac{S(v_j) * W(j)}{\sum_{t \in V(v_j)} W(t)} \quad (2)$$

$S(v_i)$和 $S(v_j)$表示节点 v_i和节点 v_j的 wpr 值,$V(v_i)$和 $V(v_j)$分别表示指向 v_i 和 v_j 的节点集合,ϕ 是阻尼系数,$W(x)$表示指向 v_x的边的权重。

2.3　查询相关性计算方法

文档 d 和查询 q 的相似度 $R(d,q)$计算公式的基本形式见公式(3)。

$$R(d,q) \approx \sum_{t \in q} w(t,q) w(t,d) \quad (3)$$

其中,t 表示词项,$w(t, q)$表示 t 在 q 中的权重,$w(t, d)$表示 t 在 d 中的权重。给定查询 q 的情况下,$w(t, q)$对结果排序没有意义,可以略去。公

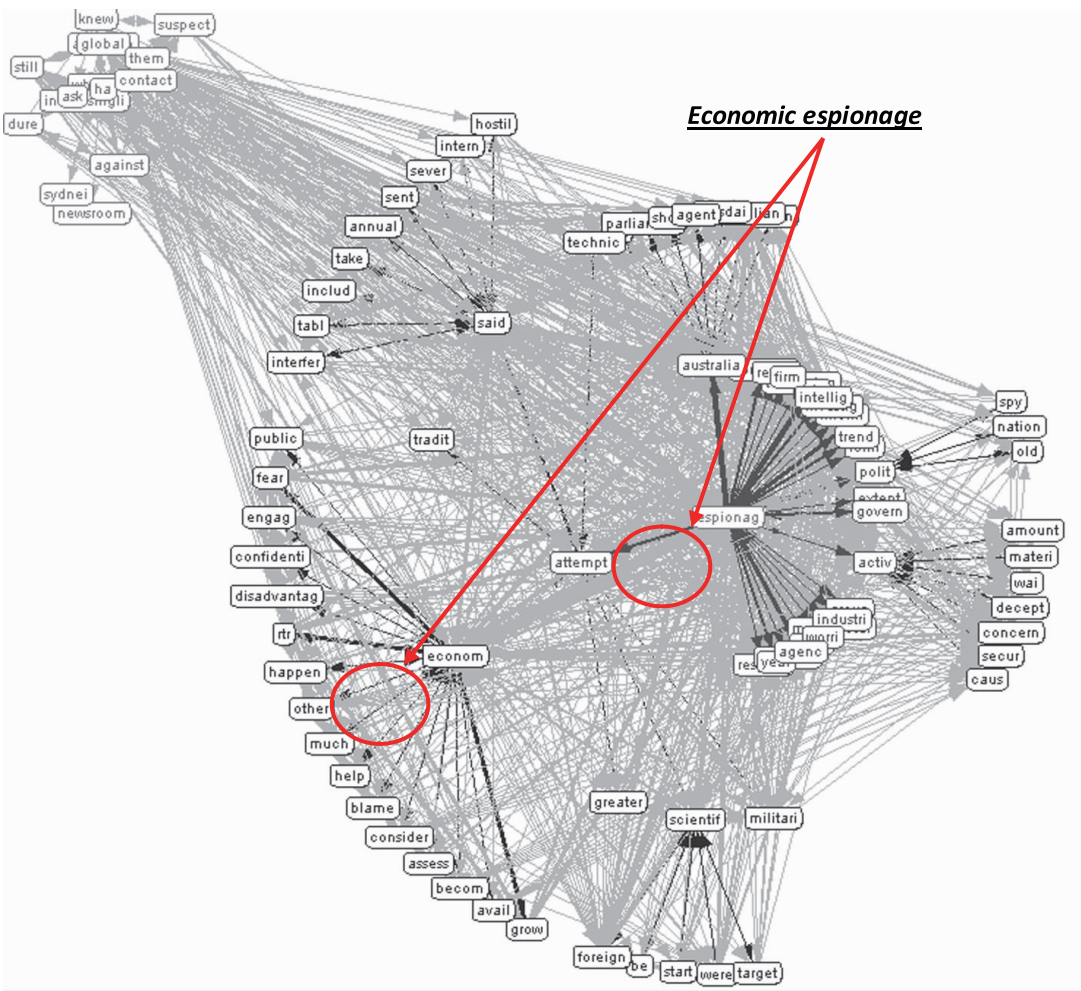

图 2　基于窗口的词性相关共词网络(依据 RCV1_250662 号文档绘制)

式为：

$$R(d,q) \approx \sum_{t \in q} w(t,d) \quad (4)$$

$W(t, d)$ 可以通过多种方法得到，在 Robertson[15] 中，使用 $w_t = \log \frac{N}{N_t}$ 作为 t 的权重，更为常见的的情况是令 $w(t,d) = freq(t) * \log \frac{N}{N_t}$，也即 $TF \cdot IDF$。IDF 在现代信息检索研究中一般被理解为词汇的信息量，freq(t)则表示词项在文档中的重要性。

调整词项重要性的度量方法，使用词项的加权 PR 值替代此处的 freq(t)，得到公式(5)。

$$R(d,q) = \sum_{t \in q} wpr(t) \log \frac{N}{N_t} \quad (5)$$

$wpr(t)$ 表示词项 t 对应节点的加权 PR 值。

为了调节 $wpr(t)$ 对文档重要性的影响，惩罚 wpr 值过高的词汇，进一步的，对 $wpr(t)$ 做 log 处理，得到基于加权网络的信息检索模型的相似度计算基本公式：

$$R(d,q) = \sum_{t \in q} \log(wpr(t)) \log \frac{N}{N_t} \quad (6)$$

同经典的概率检索模型相比，公式(6)没有利用到 tf，也没有明显的文档长度归一化因子的存在。但是，PR 值同 tf 有着一定的相关关系：高 tf 值的词汇节点一般也有着较高的 PR 值。共词网络节点规模越大，网络节点的平均加权 PR 值 $wpr(t)$ 越小，在这个意义上，公式(6)已经有了一定的归一化效果。

对应于 2.1 节提出的四种文档网络表示形式，本文得到检索模型，见表 1。

3 实 验

3.1 实验数据和方法

实验使用 Reuter RCV1 作为实验数据集。RCV1 数据集包括 806 791 篇英文文档，由路透社 1996 年 8 月 20 日到 1997 年 8 月 19 日之间的新闻文档构成，文档以 xml 格式记录，每篇文档存储为单独的 xml 文件。TREC2002 的 Filtering 任务使用 1996 年 10 月 1 日之后的文档作为评测数据集，本文的实验也仅使用这部分文档作为测试数据。实验的查询集由 TREC2002 filtering 任务的前五十个查询构成，编号为 R101 到 R150。本文使用查询的 title 域内容作为查询式。实验数据的一些参数见表 2。

在用于检索之前，文档经过了一些预处理步骤，包括 XML 解析、词干提取等。XML 解析提取文档的“title”和“text”两个节点的文本作为文档内容，不区分标题和正文。实验中分别测试了 Porter 和 Krovetz 两种 stem 模式下模型的效果。Porter stemmer 对单词词形做了较大的调整，而 Krovetz stemmer 对词的调整则相对宽松[16, 17]。实验测试了去除停用词和不去除停用词两种情况下模型的效果，停用词表为 Manning [18] 给出的 RCV1 停用词列表。基于句子的共词网络构建要求对文本进行句子切分，实验使用了 Open NLP 工具[19]。实验使用 Stanford PosTagger 对文本做词性标注[20]。构建基于窗口平移的网络时，窗口长度设定为 5。

表 1 模型对应的文档形式和权值计算方式一览表

模型	文档表示形式	相关性计算方法
Win_Weighted_Textrank	基于窗口平移的词性无关加权共词网络	公式(6)
Sent_Weighted_Textrank	基于句子的词性无关加权共词网络	
Win_Weighted_Posrank	基于窗口平移的词性相关加权共词网络	
Sent_Weighted_Posrank	基于句子的词性相关加权共词网络	

表 2 实验数据详细参数

描述	文档数量	文档集大小	词项数	查询集数量
路透数据集 disk1 + disk2 (19961001 – 19970819)	723 141	3.03GB	384 368	50

实验使用 rerank 方法验证模型的检索效果:针对 Porter 和 Krovetz 两种词干提取方法分别构建文档索引,使用 Indri 的默认模型和默认参数,取得每个查询的前三千条结果作为基准结果集[21]。实验使用 Lemur IREval[22]作为评测工具,重点关注 P@10、NDCG、Recip_rank 三个指标。由于实验目的是扩展 Blanco 提出的检索方案,本文并没有将所提出的模型同 BM25 和 Language model 做比较。

3.2 实验结果

表 3 给出了在 Porter stem 模式下各个模型的结果,表 4 给出 Krovetz stem 模式下各个模型的结果。其中,Textrank 和 Posrank 是 Blanco[11]提出的两种模型的实现:Textrank 利用词性无关共词网络表示文档,而 Posrank 使用词性相关共词网络表示文档。

从表 3 和表 4 可以看出,基于加权网络的信息检索模型较 Textrank、Posrank 有着较大的效果提升。表 3 中,Sent_Weighted_Posrank 的方法得到的检索效果最好,几乎在所有指标上都取得了最好的效果。在做停用词处理和不做停用词处理两种情况下,Sent_Weighted_Posrank 同 Posrank 相比 P@10 指标分别有着 3.1% 和 1.14% 的提升;Sent_Weighted_TextRank 相对于 Textrank 在 P@10 指标上则分别有着 3.56% 和 4.05% 的显著提升。表 4 中,Sent_weighted_Posrank 和 Win_Weighted_TextRank 都有着很好的效果,在做停用词处理的情况下,Sent_weighted Posrank 在 P@10 指标上有着 6% 的显著提升。

表 5 给出了加权方案对应于不加权方案的效果提升比例。

从表 5 可以看到,加权方案在大多数情况下较不加权方案有着显著的效果提升,这说明加权方案在本检索实验中的有效性。

表 3 Porter_stem 模式下各方法结果

Stemming	porter					
Stopwords	stopwords			nostopwords		
Performance	p@10	ndcg	recip_rank	p@10	ndcg	recip_rank
Textrank	0.5060	0.5930	0.6890	0.4940	0.5911	0.7029
Posrank	0.5160	0.5945	0.7065	0.5260	0.5989	0.7702
Win_Weighted_Textrank	0.5160	0.5970	0.7160	0.5020	0.5960	0.7280
Win_Weighted_Posrank	0.5220	0.5958	0.7213	0.5320	0.5986	0.7467
Sent_Weighted_Textrank	0.5240	0.5954	0.7034	0.5140	0.5933	0.6995
Sent_Weighted_Posrank	0.5320	0.6019	0.7587	0.5320	0.6024	0.7618

表 4 krovetz_stem 模式下各方法结果

Stemming	krovetz					
Stopwords	stopwords			nostopwords		
Performance	p@10	ndcg	recip_rank	p@10	ndcg	recip_rank
Textrank	0.5040	0.5214	0.7082	0.4840	0.5199	0.7218
Posrank	0.5000	0.5209	0.6879	0.5140	0.5233	0.7410
Win_Weighted_Textrank	0.4980	0.5255	0.7283	0.4980	0.5255	0.7283
Win_Weighted_Posrank	0.4980	0.5254	0.7281	0.4980	0.5255	0.7282
Sent_Weighted_Textrank	0.5100	0.5217	0.7088	0.4980	0.5209	0.7240
Sent_Weighted_Posrank	0.5300	0.5245	0.7207	0.5280	0.5246	0.7240

表 5　各模型较对应原始模型的效果提升比例

	Stopwords	*stopwords*			*nostopwords*		
	Performance	*p*@10	*ndcg*	*recip_rank*	*p*@10	*ndcg*	*recip_rank*
Porter	Win_Weighted_Textrank	1.98%	0.67%	3.92%	1.62%	0.83%	3.57%
	Win_Weighted_Posrank	1.16%	0.22%	2.09%	1.14%	-0.05%	-3.05%
	Sent_Weighted_Textrank	3.56%	0.40%	2.09%	4.05%	0.37%	-0.48%
	Sent_Weighted_Posrank	3.10%	1.24%	7.39%	1.14%	0.58%	-1.09%
Krovetz	Win_Weighted_Textrank	-1.19%	0.79%	2.84%	2.89%	1.08%	0.90%
	Win_Weighted_Posrank	-0.40%	0.86%	5.84%	-3.11%	0.42%	-1.73%
	Sent_Weighted_Textrank	1.19%	0.06%	0.08%	2.89%	0.19%	0.30%
	Sent_Weighted_Posrank	6.00%	0.69%	4.77%	2.72%	0.25%	-2.29%

使用词性相关网络的模型效果要好于使用词性无关网络的模型，同等实验环境下前者较后者在 P@10 指标上通常有一到三个百分点的提升。词干提取方法对模型的效果也有着一定的影响，对比表 3 和表 4，使用 Porter 的情况下，模型能够取得优于使用 Krovetz 的的效果。Sent_Weighted_Posrank 在 Porter stem 和 Krovetz stem 两种环境下取得的效果比较接近，但其他模型使用 Porter 时取得的效果要远好于使用 Krovetz。对使用和不使用停用词处理两种情况下的模型效果进行对比，可以看到停用词处理对模型的检索效果影响不大。对句子窗口和固定窗口平移两种方案进行比较，句子窗口方案明显要优于固定窗口平移方案，在同样的的实验条件下，句子窗口方案对比于固定窗口平移方案都有或多或少的效果提升。另外，加权方法在窗口平移方案构建的网络上表现相对而言不是很好，其原因还有待于进一步研究。

4　总　结

本文在 Blanco 等[10, 11]工作的基础上提出了一种基于加权网络的信息检索模型。在这一模型中，文档被表示为一个加权共词网络，词汇在文档中的重要性通过词项节点在网络中的重要性加以衡量，本文中，词项节点重要性的度量指标为 Textrank 值。本文的工作主要包括两个部分，一是使用加权网络替代无权共词网络文档表示方法，二是提出了基于句子窗口的共词网络构建方式。将加权网络应用于文档表示，可以更好的体现词汇之间的相互影响，反映词汇在整体文档中的重要性；由于句子的语言学意义，利用句子作为窗口构造共词网络相较于定长窗口平移也可以更好的反映文档的原始语义信息。本文还测试了基于词共现和基于词性两种词项关系设定方案。基于这些设定，本文提出了 Textrank 模型的四个变种，分别是 Win_Weighted_Textrank、Sent_Weighted_Textrank、Win_Weighted_Posrank、Sent_Weighted_Posrank。在 Reuter RCV1 数据集上的实验表明，本文提出的 Textrank 信息检索模型变种有着显著优于无权网络模型的检索效果。

使用文档的词共现网络表示突破了传统的词袋模型，更多地体现了词在文档中的结构信息。尽管我们的方案并没有取得超越 BM25 和 Language Model 的检索效果，但这也是在没有考虑长度归一化和词分布等信息的情况下得到的。作为对传统的词袋模型的超越，本文的研究具有一定的创新价值，后续研究中也将进一步探索文档长度、词网络拓扑性质等对检索效果的影响。

参　考　文　献

[1] Salton G. Automatic Information Organization and Retrieval [M]. New York: McGraw Hill Text, 1968.

[2] Robertson S E, van Rijsbergen C J, Porter M F. Probabilistic models of indexing and searching: Research and Development in Information Retrieval, Cambridge, 1980 [C]. Cambridge University Press.

[3] Turtle H, Croft W B. Inference networks for document retrieval: Research and Development in Information Retrieval-SIGIR, 1989[C]. ACM Press.

[4] Lafferty J, Zhai C. Document language models, query models, and risk minimization for information retrieval: Research and Development in Information Retrieval-SIGIR, 2001[C]. ACM Press.

[5] Ponte J M, Croft W B. A language modeling approach to information retrieval: Research and development in information retrieval-SIGIR, 1998[C]. ACM Press.

[6] Minsky M L. Semantic information processing[M]. Massachusetts: The MIT Press, 1969.

[7] Van Rijsbergen C J. A non-classical logic for information retrieval[J]. The Computer Journal, 1986, 29(6): 481-485.

[8] Thammasut D, Sornil O. A Graph-Based Information Retrieval System: Communications and Information Technologies, 2006. ISCIT 06. International Symposium on, 2006[C]. IEEE Press.

[9] Mihalcea R, Tarau P. TextRank: Bringing order into texts: Proceedings of EMNLP, 2004[C]. ACM Press.

[10] Blanco R, Lioma C. Graph-based term weighting for information retrieval[J]. Information Retrieval, 2012: 1-39.

[11] Blanco R, Lioma C. Random walk term weighting for information retrieval: Research and development in information retrieval-SIGIR, New York, NY, USA, 2007[C]. ACM Press.

[12] Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements[J]. Advances in Physics, 2007, 56(1): 167-242.

[13] Chomsky N. Syntactic Structures[M]. Berlin: Walter de Gruyter, 2002.

[14] Jespersen O. The philosophy of grammar[M]. Chicago: University of Chicago Press, 1992.

[15] Robertson S E, Walker S. On relevance weights with little relevance information: SIGIR, 1997[C]. ACM Press.

[16] Porter Stemming Algorithm[EB/OL]. [2012-04-22]. http://tartarus.org/~martin/PorterStemmer/index-old.html.

[17] Krovetz R. Viewing morphology as an inference process: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 1993[C]. ACM Press.

[18] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2008.

[19] Apache OpenNLP-Welcome to Apache OpenNLP[EB/OL]. [2012-04-22]. http://opennlp.apache.org/.

[20] Toutanova K, Manning C D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger: Empirical methods in natural language processing and very large corpora-SIGDAT, 2000[C]. ACM Press.

[21] INDRI-Language modeling meets inference networks [EB/OL]. [2012-04-22]. http://www.lemurproject.org/indri/.

[22] Lemur Project Home[EB/OL]. [2012-10-01]. http://www.lemurproject.org/.

(责任编辑 车 尧)