**文章编号：**

# 一种抽取微博关键短语的网络图模型 *

黄河燕，廖黎姿，王亚珅，魏骁驰

（北京理工大学 计算机学院，北京 10081）

**摘要：**微博即微博客，其简单快捷的操作方式和随时随地发布信息的互动形式使得微博内容具有噪声大、篇幅短的特点。传统的关键短语抽取方法主要运用于正式文本。当其运用于微博时，效果往往不佳。通过大量分析微博集合，发现微博用户常使用不同短语来表示相同的语义，并且这些短语的语境具有一定的相似性。提出了基于相似特征的主题化网络图模型。实验结果表明，该模型抽取微博关键短语十分高效，尤其是抽取那些被众多表达形式淹没的关键短语。

**关键词：**关键短语抽取；语境相似；网络图模型

**中图分类号：**TP391　　**文献标识码：**A

# A Graph Model for Microblog Keyphrases Extraction

HUANG Heyan, LIAO Lizi ,WANG Yashen, WEI Xiaochi

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

**Abstract:** Microblog is a new social network with the simple and quick operation for a post anytime and anywhere through the interaction form, which makes the contents quite noisy and short in length. Most existing keyphrase extraction algorithms focused on formal domains. When applied to microblog, the performance of those algorithms drops sharply. We analyze a large number of microblogs and find that people use various phrases to express the same thing while context of these phrases show similarity relationships. We propose a similarity features based topical graph model for keyphrase ranking. We evaluate our proposed model on a large microblog dataset. Experiments show that our system is very effective for keyphrase extraction, especially for digging out those keyphrases which are submerged in various forms.

**Key words:** keyphrase extraction; context similarity; graph model

## 1 引言

微博是一种允许用户即时更新简短文本(通常少于140字)并且可以公开发布的微型博客形式[1]。中文微博在近两年迅速发展，以新浪微博为代表，包括腾讯、搜狐、网易、凤凰等其他门户纷纷加入微博阵营。中国互联网信息中心（CNNIC）数据显示，至2012年6月底，中国微博的用户总数达到2.74亿，成为微博用户世界第一大国[2]。尽管微博内容冗杂，形式不正式，但却能提供有关微博用户想法和观点的第一手信息。由于关键短语可以很好地概括微博内容，因而，从这些实时的海量信息中抽取关键短语将对热点发现、舆情监控、危机公关等有积极作用。

总的来说，完成关键短语抽取的方法可大致分为两类。一类是将关键短语抽取当成二分类任务来解决，另一类则是将关键短语抽取当成排序任务来解决。在二分类任务中，文本中的每个词语被标记为关键短语或非关键短语。Turney[3, 4] 将基于频率和 POS 的信息作为特征，运用遗传算法来调节一系列参数化的启发式规则，提出在文档关键短语抽取方面取得进

* **收稿日期：** 2013/7/7　　**定稿日期：**2013/7/13

**基金项目：**国家重点基础研究发展计划（973计划）(2013CB329300)

**作者简介：黄河燕**（1963——），女，教授，主要研究方向为信息抽取、机器翻译；廖黎姿（1989——），女，博士研究生，主要研究方向为信息抽取；王亚珅（1989——），男，博士研究生，主要研究方向为社会计算；魏骁驰（1990——），男，博士研究生，主要研究方向为信息检索、数据挖掘。

展的 GenEx。Frank 等开发了系统 KEA[5, 6]，运用 TFIDF 和首现位置作为特征，采用朴素贝叶斯技术对短语离散的特征值进行训练，获取模型的权值。Hulth 探索了更多的语言学知识[7]。Medelyan 和 Witten 通过从领域特殊化的词库中提取词和短语的语义信息以提升关键短语自动抽取的表现[8]。在非监督的排序任务中，根据不同的特征给每个候选短语打分，然后排在最前面的小部分候选短语被选为关键短语。其中，基于网络图模型的排序方法是当前效果最好的[9,10]。首先，这些方法运用文本内部的某些关系(比如词共现)来建立网络图。然后，用随机漫步技术来计算词的重要度。最后，排在最前面的词用于选取关键短语。考虑到主题信息，Liu 等把传统的 PageRank 技术分解成主题化的 PageRank[9]。Zhao 等又将主题化的 PageRank 扩展成语境敏感的主题化 PageRank，在排序时考虑相关度和兴趣度信息[10]。这些方法都采用关键短语抽取的标准三步法，即关键词排序、候选关键短语生成、关键短语排序。其中，候选关键短语由排在前面的关键词组合而成，并且要求在数据集中频繁出现。

然而，这种标准三步法运用于微博关键短语抽取时，存在一定的弊端。微博的内容十分混杂，用户常用不同的短语来表达同一语义。海量微博由数以亿计的微博用户写成，他们的文化背景，立场等存在较大差异。他们看待事物的观点、态度等往往会有不同，进而语义相同的微博可能会使用不同的词语。从这个角度来看，用关键词来组合候选关键短语的方法不能完全囊括各种不同的表达形式，语义相同但用词不同的短语不应该被区别对待。

我们通过分析大量微博集合发现，当一个事件或话题出现时，人们常使用不同的词语组合成短语来进行表述，从而，关键短语往往会有多样的形式。同时，我们发现虽然这些短语的形式多样，但是其语境存在相似关系。因此，我们提出基于相似特征的主题化网络图模型。首先利用相似特征将指代同一语义单元的短语单元合并成一个集合，然后将集合作为节点，通过共现关系建立网络图模型对节点进行排序。从而获取关键短语。

文章首先分析关键短语的多样形式。然后重点介绍基于该现象我们提出的主题化微博关键短语抽取方法。随后，对模型进行参数估计以及实验比较。最后得到研究结论并且提出下一步工作计划。

## 2 关键短语的多样形式

由于没有一个统一的标准来区别不同的短语是否表示同一个信息单元，我们采用 Van Halteren[11]以及 VahedQazvinian[12]提出的方法。本文定义短语单元和信息单元。每个候选短语都被视为独立的短语单元，而不同的短语单元可以代表同一个信息单元。

被分析的微博均摘自 weibo.com[13]，由不同的人写成。随机阅读 n 条微博，计算所包含的短语单元和信息单元个数。这些数字代表被阅读微博传达的信息量。通过这个实验，我们可以得到关于微博多样性的一些启示。图 1 表示每一个 n 所对应观察到的不同短语单元和信息单元数目。曲线用对数坐标描绘，以突出短语单元数目与信息单元数目变化的差异。

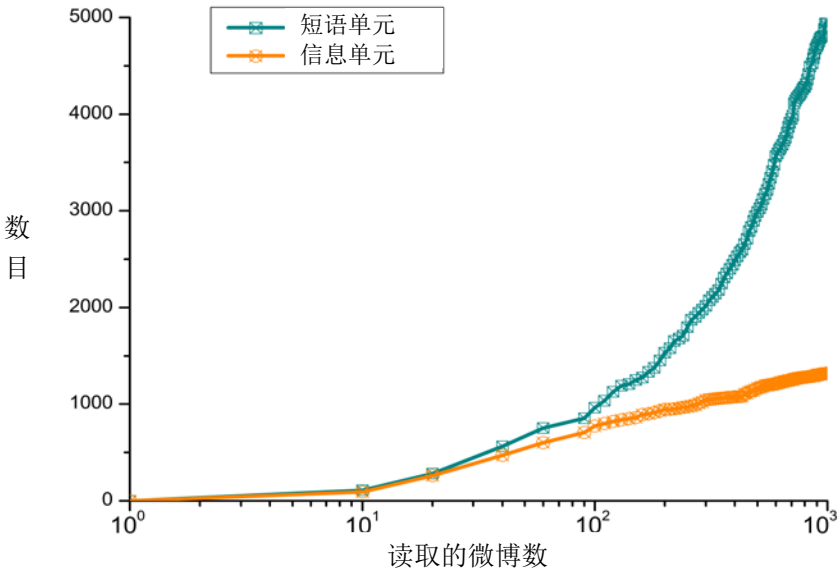


图 1　随机读取 n 条微博所观察到的不同短语单元和信息单元数量

由图 1 中可见，随着被观察微博数目的增长，观察到的短语单元和信息单元数目都在增长。虽然短语单元数目和信息单元数目的增长都没有表现出某种渐进趋势，但是可以观察到在后期短语单元数目的增长速度明显大于信息单元数目的增长速度。在初期，被观察的微博数目少，信息单元重叠概率较小，从而数量分布与短语单元近似。但是随着被观察微博数目增多，信息单元重叠概率加大，信息单元数目的增长速度开始逐步落后于短语单元数目的增长速度。从这我们可以得出一个结论，即微博关键短语的多样性一定程度源于用词的多样性。

接下来，通过实例，我们分析一个有趣的现象。如下是关于神舟九号与天空一号对接成功的微博，两条微博中的短语单元由人工标出，用下划线表示：

M1：祝贺 中国航天！神舟九号， 与天宫一号，手动对接 成功。

M2：6 月 24 日，神九与天宫一号 牵手。满分！

可以看出，在这两条微博中总共有 6 个信息单元由 11 个短语单元表示。

f1：{神舟九号，神九}

f2：{天宫一号，天宫一号}

f3：{手动对接，牵手}

f4：{祝贺，成功，满分}

f5：{6 月 24 日}

f6：{中国航天}

通过分析可以发现，当一个事件发生或者一个话题出现时，人们往往用不同的短语单元来写出不同的微博，尽管主题内容可能是一致的(比如，‘神舟九号’与‘神九’)。从这个意义上来说，这些不同的短语单元被用来指代相同的信息单元。

进一步来看，对于微博 M1 和 M2，我们可以分别用一个短语集合来表示‘神舟九号’和‘神九’。其中，短语集合中的短语均取自被表示短语的周围。

神舟九号：{天宫一号，手动对接，成功，祝贺，中国航天}

神九：{天宫一号，牵手，满分，6 月 24 日}

以上展示出一个有趣的现象，尽管‘神舟九号’和‘神九’在形式上不同，但是集合表示却十分近似。这种现象我们可以在很多情况中观察到，比如一个新电影被播出，一个社会问题被曝光，一个重大科学发现被报道等。人们总是尝试着从不同的角度用不同的短语单元来描述同一信息单元。然而，尽管这些短语单元在形式上展现出某些的差异，表示他们的集

合却往往向我们展示出一定程度的相似性。

## 3 关键短语抽取研究方案

### 3．1 主题化分解

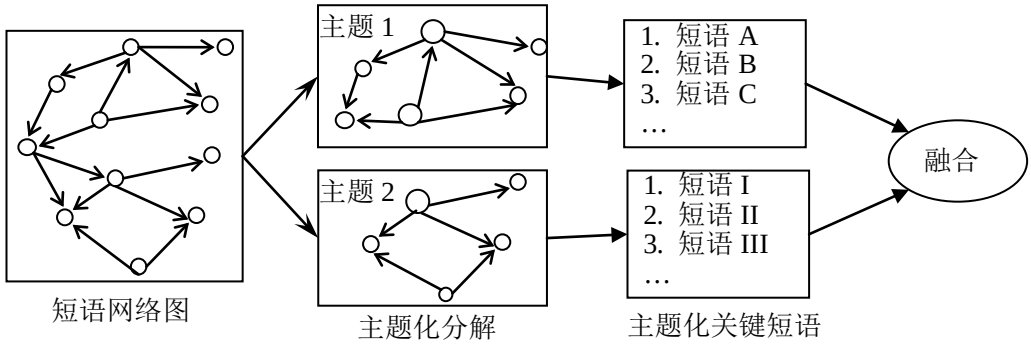


图 2　关键短语抽取的主题化分解

图 2 对微博集合主题化分解进行了简单示意。主题化分解有两个好处。一方面，微博集合被分解成相对较小的几个模块，后续处理的计算规模大大减小。另一方面，关于具体某个主题的关键短语更有针对性，掺杂的噪声较少，更具有研究意义。

微博内容有长度限制，并且每条微博往往都是关于一个单一主题。Weng 等人证明作者主题模型对微博进行主题建模十分高效[14]。假设每条单一微博只有唯一的主题，采用修正的作者主题模型进行建模。该模型为微博 LDA。该模型假设整个微博集合中有一系列主题，这些主题被表示表示为词语的分布。同时，每个作者都有其自身的主题偏好，该偏好被建模为主题的分布。在写微博时，作者根据其主题分布选择一个主题，然后根据该主题依次选择词语组成微博。

### 3．2 网络图模型

Hulth 曾经指出手动标注的关键短语主要都是名词短语[15]。从而，我们只从微博中选取基本名词短语作为候选短语。所谓基本名词短语是指由一个核心名词以及不超过 3 个前置形容词或名词修饰成分组成的短语，其中不含递归结构。另外，由于微博有一个特点，即用户可以自己赋一个主题并写在两个’#’之间。考虑到这一特性，我们把人为赋予的主题也当做候选短语。为获取微博中的候选短语，我们对微博进行词性标注(POS)，并且删除其中的 stop 词。然后，我们用类似 Liu 工作[9]中的(adjective)*(noun)+模式来抽取基本名词短语，这个模式表示在一个或多个形容词后跟随着一个或多个名词。同时，我们要求组成基本名词短语的形容词或名词个数总和不超过 4。

Mihalcea[16]曾指出，网络图模型的方向对关键短语抽取的影响不大。因而，我们建立带有方向的网络图。定 $G=(V,E)$ 为微博数据集所对应的网络图。其中有节点集合 $V=\{p_1, p_2, \cdots, p_n\}$，并且设定边集合 $(p_i,p_j)\in E$ 。设定边的权重如式:

$$e(p_i,p_j)=context\_sim(p_i,p_j)+\alpha*headnoun\_sim(p_i,p_j) \quad (1)$$

其中，$context\_sim(p_i, p_j)$ 为语境相似度值，$headnoun\_sim(p_i, p_j)$ 为核心名词相似度值。同时，节点 $p_i$ 的度设置为按公式(2)求取:

$$O(p_i)=\sum_{j:p_i\to p_j} e(p_i,p_j) \quad (2)$$

3．2．1 语境相似关系

考虑到微博用词的多样性，本文用候选短语的语境相似关系来解决同一信息单元被不同的短语单元指代的问题。每一个候选短语都用其上下文集合来表示。候选短语的语境相似度则通过计算相应表示集合的相似度给出。通过这种方式，我们间接引入了短语句法关系。句

法关系假设一定距离内的词语相互之间具有句法或语义上的联系。Lyon[17]指出，约 70%的句法关系依赖于直接邻居，约 17%的句法关系依赖距离为 2。在我们的方法中，每一个候选短语 $p_i$ 都由其最近邻的 6 个(或少于)词或短语组成的集合表示。通过计算候选短语对($p_i$，$p_j$)所对应的集合 $l_i$ 与 $l_j$ 的相似度，我们得到两候选短语对的语境相似度。集合相似通过 HowNet[18]的概念近似获取，即使用其提供的 API 函数:HowNet_Get_Concept_Similarity。

3．2．2 核心名词相似关系

本文中用候选短语中核心名词的相似关系作为内部的相似关系。这种做法主要有两个原因：首先，带有更多修饰词的名词短语往往更为特殊化，更为紧密地联系着一个主题或事物的某一特殊方面。然而对于关键短语抽取来说，太过具体化的信息反而缺乏代表性。比如，’红十字会’与’江苏电大红十字会’。另外，当用户关于某一话题或事物写微博时，可能因为不同的立场或感受而使用不同的修饰词。因为中文短语中的核心名词通常是位于最后的名词，我们把短语中最后一个名词作为核心名词。然后，计算两个核心名词之间的相似度值。

### 3．3 主题化关键短语排序

本文通过 PageRank 技术给候选短语赋分值排序[19]并用马尔科夫链和随机行走过程解释工作原理。 PageRank 通过网络的超链接关系来确定一个页面的等级。简单地说，一个高等级的页面可以使相连的低等级页面提升等级。类似页面间的投票或推荐。也就是说，在构建的网络图中，一个节点的重要程度(即 PageRank 值)由指向它的其他节点的 PageRank 值之和决定。由于一个节点可能指向许多其他节点，那么它的 PageRank 值将被所有它指向的节点共享。

在构建好的网络图中，设总共有 N 个节点。可以用矩阵来表示所有的等式。用 S 代表表示 PageRank 值的 N 维列向量。用 A 表示网络图的邻接矩阵，可得:

$$A_{ij}=\begin{cases}\dfrac{e(p_i,p_j)}{O(p_i)} & (p_i,p_j)\in E\\ 0 & \text{其他}\end{cases} \tag{3}$$

且满足从任意节点出发，转移到其他节点的概率和为 1，即:

$$\sum_{j=1}^{N}A_{ij}=1 \tag{4}$$

根据马尔科夫链的各态经历理论[20,21]，如果随机转移矩阵 A 是不可约且非周期的，则其定义的有限马尔科夫链具有唯一的静态概率分布，即经过一系列的状态转移后，不论每个节点的初始概率是什么，最后它们的概率值都会达到收敛。用该静态概率分布作为 PageRank 向量 S，得到:

$$S=A^T\times S \tag{5}$$

为保证转移矩阵 A 为不可约且非周期的，加入转移概率因子 λ，取值区间为(0，1)。从而，节点 $p_i$ 的 PageRank 值 $s(p_i)$ 可由公式(6)计算获得:

$$s(p_i)=\lambda\sum_{j:P_j\to p_i}\frac{e(p_j,p_i)}{O(p_j)}s(p_j)+(1-\lambda)\frac{1}{|V|} \tag{6}$$

根据以上理论，Liu 等人提出主题化的 PageRank 以确定关键短语排序中的关键词[9]。文中称 TPR。对于每一个主题，运行针对主题的 PageRank。公式如下：

$$R_t(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{e(w_j,w_i)}{O(w_j)}R_t(w_j)+(1-\lambda)P_t(w_i) \tag{7}$$

其中，$R_t(w)$ 是词语 $w$ 关于主题 $t$ 的主题化 PageRank 值。 λ 为取值 0 到 1 的转移概率因子。 $e(w_j,w_i)$ 为边$(w_j\to w_i)$ 的权值。

Zhao 等人发现 TPR 在设置边的权值时忽略了主题语境[10]。这种情况可能导致最后的 PageRank 分值失去话题针对性。从而，他们提出语境敏感的 PageRank 方法。文中称 cTPR。

$$R_t(w_i)=\lambda\sum_{j:w_j\to w_i}\frac{e_t(w_j,w_i)}{O_t(w_j)}R_t(w_j)+(1-\lambda)P_t(w_i) \tag{8}$$

其中，从 $w_j$ 指向 $w_i$ 的边权值根据主题来赋值。即网络图模型中的边亦被主题化。

然而，TPR 和 cTPR 都将短语分割成独立的词语并对词语进行排序，忽略了这种方法可能向关键短语排序掺入噪声，尤其是在微博这种表达形式多样的数据集上。微博中大量使用不同短语单元指代相同信息单元的现象在前面已有分析。从而，文章提出基于相似特征的主题化微博关键短语抽取方法。

$$R_t(S_i)=\lambda\sum_{\substack{k_n\to k_m\\ k_n\in S_j\\ k_m\in S_i}}\frac{e_t(k_n,k_m)}{\sum_{k\in S_j}O_t(k)}R_t(S_j)+(1-\lambda)P_t(S_i) \tag{9}$$

式中 $R_t(S)$ 为短语集合 $S$ 关于主题 $t$ 的基于相似特征主题化 PageRank 分值。集合 $S$ 中为根据相似特征汇聚到一起的候选短语 $k$。由于两个相似特征为语境相似度和核心名词相似度，文中称该方法为 CH。

## 4 实验结果与分析

### 4．1 数据集和预处理

由于目前还没有针对微博关键短语抽取的标准数据集，我们构建了一个数据集。收集的大部分微博是在 2011 年 6 月到 2012 年 8 月期间发表的。这些微博随机摘取自 weibo[13]。过滤掉常见 stop 词(如语气助词等)以及出现在少于 10 条微博中的词。另外过滤掉被少于 10 个用户评论或转发的微博以及作者总共发微博数少于 10 的微博。为了模拟微博中转发这一功能，当一条微博被转发时我们简单地在数据集中复制一遍。表 1 中给出数据集的一些统计信息。

表 1 数据集的一些统计信息

| #作者 | #微博数 | #候选短语数 |
|---|---|---|
| 1，570 | 32，400 | 14，750 |

### 4．2 评价指标

对于在数据集上进行的实验，使用的评价指标类似于信息检索中广泛运用的 nDCG 方法[22]。其计算公式如式(7)所示：

$$nDCG@K=\frac{\sum_{j=1}^{K}\frac{1}{\log_2(j+1)}score(j)}{IdealScore(K)} \tag{10}$$

其中，score(•)为实际获得的分值。IdealScore(K)作为归一化因子，是理想排序下 K 个关键短语的分值。如果一个方法在前排获得更多优良的关键短语，那么其 nDCG 分值越高。

### 4．3 参数估计

方法中主要有两个参数影响关键短语抽取的效果:(1)语境相似度与核心名词相似度的比例调节因子 $\alpha$，(2)PageRank 算法中的转移概率因子。在这一节中，研究这些参数对抽取效果的影响。

参数 $\alpha$ 调节语境相似度与核心名词相似度对构建网络图的影响比例。当 $\alpha>1$ 时，可以增加核心名词相似度对构建网络图产生的影响。当 $\alpha<1$ 时，语境相似度对其影响较大。实验结果如图 2 所示，$\alpha=1.5$ 时，方法表现更好。语境相似度在 0 到 6 之间离散变化，而核心名词相似度则为 0 或者 1。 由于两个特征取值区间已经存在较大差异， $\alpha$ 的值取得相对较小。当这两种特征的值都不为 0 的时候，带有权重的网络边被建立。

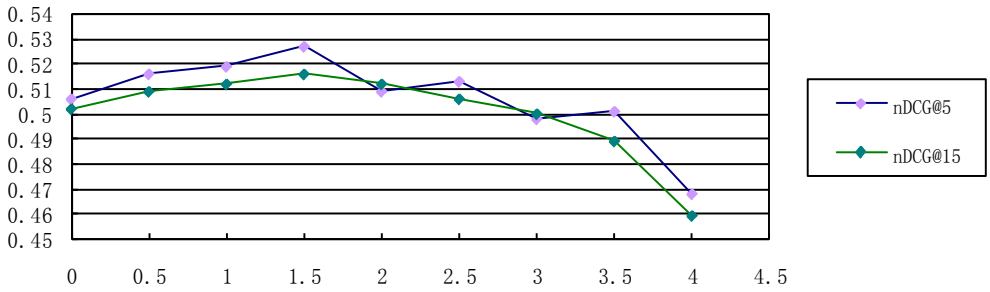


图 3 不同 $\alpha$ 值对关键短语抽取效果的影响

对于 $\lambda$，我们以 0.05 为步长，从 0.05 实验到 0.95，最后得到为 $\lambda = 0.15$ 该方法的最优设置。 转移概率因子缓和图遍历以及偏好的影响。

### 4．4 实验结果

文中方法结合两个特征，即语境相似度和核心名词相似度，故简称 CH。将该方法与目前效果最好的 cTPR 以及 TPR 作比较。实验中 2 个基线方法列表如下:

(1) TPR。该方法为清华大学的主题化 PageRank 方法，应用于正式文本关键短语抽取。

(2) cTPR。该方法为北京大学的语境敏感主题化 PageRank 方法，应用于 Twitter 的关键短语抽取，考虑 relevance 和 interestingness 信息，为目前效果最好的方法。

在数据集上分别运行各种方法，然后对结果进行过滤。删除掉无意义的短语以及重复的短语。然后对这些方法分别计算 nDCG 分值。 为了更详细地比较结果，我们设定 n 为 5 到 50 进行实验。实验结果表明，CH 方法效果总体比其他的方法好。实验结果列在图 4 中。表中的粗体表示相同的 n 取值下，微博关键短语抽取的最好结果。

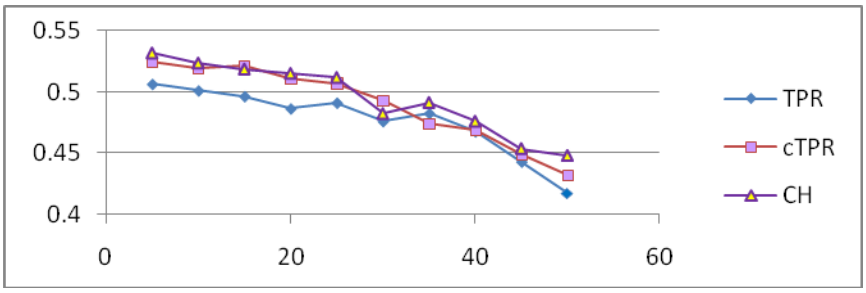


图 4　CH 方法与基线方法比较结果

由于 cTPR 方法的目标是提取主题关键短语，为了让其与本文方法的比较更为全面，根据不同的主题人工选择 50 个主题微博集合，并在这些集合上分别进行比较试验。对于 n，取值 5，10，25 以进行详细比较。因为空间的限制，表 2 仅列出一些最具代表意义的结果。

表 2 CH 方法与 cTPR 方法在人工数据集上的比较

| | | nDCG@5 | nDCG@10 | nDCG@25 |
|---|---|---|---|---|
| 神舟九号 | cTPR | 0.767 | 0.753 | 0.749 |
| | CH | **0.821** | **0.813** | **0.798** |
| 红十字会 | cTPR | 0.794 | 0.790 | **0.785** |
| | CH | **0.811** | **0.801** | 0.784 |
| … | … | … | … | … |
| 欧洲杯 | cTPR | **0.824** | **0.814** | 0.792 |
| | CH | 0.812 | 0.802 | **0.794** |

这些数据集包含噪声较少。从表中可以看出， CH 方法与 cTPR 方法的工作效果都比在整个数据集上的效果好很多。详细比较每个数据集上的表现可以发现，在神舟九号数据集上，

CH 方法的效果尤其比 cTPR 方法的效果好。神舟九号有一个广为人知的替代形式为神九。在微博中，尽管二者在形式上不同，但二者的语境是十分近似的，往往都是关于航天或者对接。CH 方法中的语境相似度特征非常适合于捕捉这一信息，并将其运用到关键短语的抽取中。在红十字会数据集上， CH 方法的效果稍微比 cTPR 方法的效果好一点。红十字会是许多名词短语的核心名词，如中国红十字会，北京红十字会等。 CH 方法中的核心名词相似度特征有效地考虑到了这种情况。

## 5 结束语

无监督的微博关键短语抽取方法，将关键短语抽取转换成在网络图上节点的排序问题。本文在已有网络图模型排序算法基础上引入相似特征，并修正标准三步法，提出了基于相似特征的微博关键短语抽取方法。实验表明，综合的两个相似特征可以提高微博关键短语抽取的性能，尤其是对那些具有较多替代形式的关键短语。

本文下一步的改进工作主要从集中于深入探索语境相似度特征和核心名词特征的表现方法。这两个特征在文中被设为离散值并且用线性方法组合。应该存在更好的方法拟合二者之间的关系。

## 参考文献


[1] http://zh.wikipedia.org/wiki/%E5%BE%AE%E5%8D%9A%E5%AE%A2

[2] http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201207/t20120723_32497.htm

[3] Peter D.Turney. Learning to extract keyphrases from text. Technical Report,NRC/ERB-1057,Ottawa: National Research Council Canada, 1999.

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

[5] Eibe Frank, Gordon W.Paynter, Ian H. Witten, Carl Gutwin, and Craig G.Nevill-Manning. Domain-specific keyphrase extraction. InProceedings of the16st International Joint Conference on Artificial Intelligence (IJCAI'99),Stockholm, Sweden: Morgan Kaufmann, 1999. 668–673.

[6] Witten, Ian H.,FordonW.Paynter,Eibe Frank, Carl Gutwin&CraigG.Necill-Manning.KEA: Practical Automatic Keyphrase Extraction. In: Neil Rowe ed. Proc. of the 4th ACM Conference on Digital Libraries.Berkeley,CA : ACM, 1999. 254-255.

[7] AnetteHulth. Improved automatic keyword extraction given more linguistic knowledge. In: Michael Collins ed. Proc. of EMNLP-2003,Sapporo: Association for Computational Linguistics, 2003. 216-223.

[8] OlenaMedelyan, Ian H. Witten. Thesaurus based automatic keyphrase indexing. InProceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, New York:ACM, 2006. 296-297.

[9] ZhiyuanLiu,WenyiHuang,YabinZheng and Maosong Sun. Automatic keyphrase extraction via topic decomposition. In: Hang Li eds.Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing.Massachusetts: Association for Computational Linguistics, 2010. 366-376.

[10] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song,PalakornAchananuparp,Ee-peng Lim and Xiaoming Li. Topical keyphrase extraction from Twitter. In: Dekang Lin ed. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Port-land, Oregon: Association for Computational Linguistics,2011.379-388.

[11] Hans Van Halteren and Simone Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis.In:Drago and Simone eds. Proc. of the HLT-NAACL 03 on Text summarization workshop. Stroudsburg: Association for Computational Linguistics,2003. 57-64.

[12] VahedQazvinian and DragomirR.Radev. Learning from collective human behavior to introduce diversity in lexical choice. In: Dekang Lin ed.Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Port-land, Oregon: Association for Computational Linguistics, 2011. 1098-1108.

[13] http://weibo.com/

[14] JianshuWeng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM International Conference on Web Search and Data Mining. 2010.

[15] AnetteHulth. Improved automatic keyword extraction given more linguistic knowledge. In*Proceedings of EMNLP*, pages 216-223, 2003.

[16] MihalceaR,TarauP.Textrank: Bringing order into texts. In Proc. of the Conf. on Empirical Methods in NaturalLanguage Processing. Morristown: Association for Computational Linguistics, 2004.404-411.

[17] Lyon, C., Nehaniv, C., and Dickerson, B.: Entropy indicators for Investigating early lan-guage process. In: AISB'05: Proc. of EELC'05, Pages 143-150 (2005).

[18] http://www.keenage.com/

[19] Page L,BrinS,MotwaniR,Winograd T. The PageRank citation ranking: Bringing order to the Web. Vol.66. Technologies Project, San Francisco: Stanford InfoLab., 1998.281-287.

[20] Liu B. Web Data Mining. Heidelberg: Springer-Verlag, 2007. 245-254.

[21] Aiello W, Chung F, Lu LY. A random graph model for massive graphs. In: Giannotti F ed. Proc. of the ACM Symp.on Theory ofComputing. Pittsburgh: Association for the Advancement of Artificial Intelligence, 2000. 171-180. http://bigcheese.math.sc.edu/~lu/papers/random.pdf

[22] KalervoJärvelin and JaanaKekäläinen. Cumulated gain-based evaluation of ir techniques. ACM Trans. on Information Systems,2002,20(4):422-466.