基于超图的文本摘要与关键词协同抽取研究*

莫鹏，胡珀，黄湘冀，何婷婷

（华中师范大学 计算机学院，湖北省武汉市 430079）

摘要：文本摘要和关键词抽取是自然语言处理领域的两个重要研究课题，它们均以生成描述文本主旨内容的精简信息为目标。尽管这两个任务目标相似，但它们通常被作为两个独立的问题分别研究，而较少考虑其彼此间的自然关联性。尽管已有学者提出了基于图模型的协同抽取方法，该方法同时考虑了句子与句子、词与词、句子与词之间的各种关系，以迭代强化的方式同时生成文本摘要和关键词，但现有模型大多仅限于表达句子与词之间的各种二元关系，而忽视了不同文本单元间潜在的若干重要的高阶关系。鉴于此，本文提出了一种新的基于超图的协同抽取方法。该方法以句子作为超边，以词作为结点构建超图，在一个统一的超图模型下同时利用句子与词之间的高阶信息来生成摘要和关键词。在 NLPCC 2015 面向微博的新闻文本摘要任务数据集上的实验结果验证了本文所提方法的可行性和有效性。

关键词：超图；文本摘要；关键词抽取；协同抽取

中图分类号：TP391　　**文献标识码：**A

A Hypergraph Based Approach for Simultaneous Text Summarization and Keyword Extraction

Peng Mo, Po Hu, Xiangji Huang, Tingting He

(Central China Normal University, School of Computer Science, Wuhan, Hubei 430079, China)

Abstract: Text summarization and keyword extraction are two important research topics in Natural Language Processing (NLP), and they both generate concise information to describe the gist of text. Although these two tasks have similar objective, they are usually studied independently and their association is less considered. Following the graph-based ranking methodology, some collaborative extraction methods have been proposed, which considered the association of sentences, words and the relationships between sentences and words, and generated both text summary and keywords in an iterative reinforced framework. However, most existing models are limited to express various kinds of binary relations between sentences and words, which ignore a number of potential important high-order relationships among different text units. Because of these, we propose a new collaborative extraction method based on hypergraph. In this method, sentences are modeled as hyperedges and words are modeled as vertices to build a hypergraph, and then summary and keywords are generated by taking advantage of higher order information from sentences and words under the unified hypergraph. Experiments conducted on the Weibo-oriented Chinese news summarization task in NLPCC 2015 demonstrate that the proposed method is feasible and effective.

Key words: hypergraph; document Summarization; keyword extraction; collaborative extraction

1 引言

文本摘要是从给定的文本中生成能够表达原文主题的精简摘要;关键词抽取是从给定文章中抽取出重要的词或短语，这些词或短语能够代表原文的重要信息。上述两种研究任务有一定的相似性，它们的目的都是获取文章的简洁表达。文本摘要和关键词抽取之所以受到很大关注，是因为它们在文本挖掘的很多领域有很重要的应用，包括文本检索、文本聚类等。比如文本摘要能够方便用户快速浏览搜索的结果并且快速找到所需信息;而关键词可以作为

* **收稿日期：2015-06-18**　　**定稿日期：2015-08-09**

基金项目：国家自然科学基金青年科学基金项目 （61402191）、华中师范大学中央高校基本科研业务费项目（CCNU14A05015、CCNU15ZD003）、国家社科基金重大计划招标项目（12&2D223）、华中师范大学教师科研启动基金项目

文章的索引，以此来提高文本检索的准确性。

文本摘要和关键词抽取都可以分为一般式的和查询相关的。一般式的文本摘要和关键词抽取要尽可能覆盖原文的重要信息，它面向的是所有用户。而查询相关的文本摘要和关键词抽取是在给定一个查询条件下，生成的摘要和抽取出的关键词要尽可能地和这个查询相关，它面向的是特定用户需求。还可按照处理文本的数量将其划分为单文档式和多文档式。本文关注的是一般式的单文档摘要和关键词抽取。

文本摘要和关键词抽取已经在自然语言处理和信息检索领域被广泛研究。近年来，基于图的排序算法已经成功地应用于文本摘要[1][2]和关键词抽取[1]。通常，这些方法利用投票机制从文本中抽取出句子或关键词。虽然这两个任务有很多共同之处，但是大多数方法仅研究其中一个任务。

Wan 等人提出了一种基于图的迭代强化方法[3]在一个框架下同时处理这两个任务，该方法基于两个假设，假设 1：一个句子如果与其他重要的句子有密切的关系，那么这个句子也是重要的；一个词如果与其它重要的词关系密切，则这个词也是重要的。假设 2：一个句子如果包含许多重要的词，那么这个句子应该是重要的；一个词如果出现在许多重要的句子中，则这个词也是重要的。该方法同时考虑了句子与词之间的三种关系（句子与句子之间的同质关系，词与词之间的同质关系，句子与词之间的异质关系），分别构造了三种关系图（SS-Graph、WW-Graph、SW-Graph），句子和词之间的上述三种关系相互强化，迭代计算出句子和词的重要性得分，直到收敛。此方法仅能表达句子与词之间的各种二元关系，而忽视了不同文本单元间潜在的若干重要的高阶关系。为此，本文提出了一种基于超图的文本摘要和关键词协同抽取方法 HBCE（Hypergraph-based Collaborative Extraction），以句子作为超边，以词作为结点构建超图，在一个统一的超图模型下同时利用句子与词之间的高阶信息来生成摘要和关键词。另外，该方法不需要利用知识库或背景语料，使得该方法在模型和计算复杂度上都更有优势。

2 相关工作

2.1 文本摘要

文本摘要的方法通常分为抽取式和摘要式，本文主要研究抽取式单文本摘要。抽取式文本摘要首先给文章中的每个句子打分，根据分数将句子排序。句子分数的计算通常结合了统计学和语言学特征，包括词频、句子位置特征、线索词和主题标签[4][5]等。也有一些利用机器学习抽取句子的方法，有非监督的方法[6]和监督的方法[7][8]，除此之外，还有最大边际相关性（MMR）[9]、潜在语义分析（LSA）[10]等方法。基于图的方法也被广泛应用于抽取式文本摘要任务中，包括 TextRank[1]和 LexPageRank[2]，这些方法基于句子相似性，构建关系矩阵，每个句子的重要性由和它相关的句子来决定。这些方法通常都只考虑了句子和句子之间的关系。Wan 等人提出了一种基于图的迭代强化方法[3]，该方法同时考虑了句子与句子、词与词、句子与词之间的各种关系，在句子排序的过程中融入了词的强化作用。

2.2 关键词抽取

传统的关键词抽取方法仅仅使用文本中包含的显式信息，如词频和位置等。Salton 和 Buckley 提出了一个简单的基于词频的关键词抽取方法[11]。基于图的方法有 TextRank[1]，该方法使用了三种统计属性信息，包括 $tf \times idf$ 、距离和关键短语频率。还有 Wan 等人提出的基于图的迭代强化方法[3]，考虑了句子对词的强化作用。Ercan and Cicekli 提出了使用词汇

链特征的方法[12]。比较流行的还有基于机器学习的关键词抽取方法[13][14][15]，通过机器学习算法为候选词分类，进而判断候选词是否属于关键词。目前，已经有很多方法开始使用 Wikipedia 作为背景知识来抽取文本中的关键词[16][17]。

3 基于超图的协同抽取方法

3.1 构造超图

定义 $HG(V,E,w)$ 为一个有权无向超图，点集合为 V，边集合为 E。超边 e 是 V 的一个子集 $\cup_{e\in E} e = V$，e 的权重为 $w:\mathcal{E}\to$ ¡ $^{|E|}_{+}$。如果 $v\in e$，则超边 e 与 v 关联。超图的关联矩阵定义为 $H\in$ ¡ $^{|V|\times|E|}$ 如下：

$$h(v,e)=\begin{cases}1 & \text{if } v\in e\\ 0 & \text{if } v\notin e\end{cases} \tag{1}$$

点和超边的度定义如下：

$$d(v)=\sum_{e\in E} w(e)h(v,e) \tag{2}$$

$$\delta(e)=\sum_{v\in V} h(v,e)=|e| \tag{3}$$

D_e 和 D_v 分别代表超边和点的度对角矩阵。W_e 是边权重对角矩阵。

3.2 超边权重计算

超图里的每一条边代表一个句子，句子的权重可以由这个句子的主题信息密度来衡量。对于词 $w\in V$，假设词 w 所包含的主题信息为 $T(w)$，则句子的主题信息密度的计算方式如下：

$$w(e_i)=\frac{1+\sum_{w\in e_i} T(w)}{\delta(e_i)^2} \tag{4}$$

$\sum_{w\in e_i} T(w)$ 代表句子 e_i 包含的所有词主题信息之和，由于包含的词越多，包含的主题信息越多，而长句相比短句会有更大的概率包含更多的主题信息，所以我们采用 $\delta(e_i)^2$，即句子长度的平方来有效地惩罚长句。

3.3 基于边的随机游走

我们可以把漫游者在超图上的随机游走过程看作是在传统图上游走的一个推广，在传统的图上，漫游者沿着边的游走可以看作是选择与该边相连的点，然后转移到另外一条边上。但是，在超图上一条边可能与多个点相连，因此我们需要一般化这个过程。漫游者在超图上游走分为两步：第一步，漫游者从当前所在超边 e_i 上随机选择一个点 v；第二步，漫游者以 $w(e_j)$ 与包含 v 的所有边的权重之和 $\sum_{\hat{e}\in E(v)} w(\hat{e})$ 的比值作为概率选择边 e_j，且满足 $v\in e_i\cap e_j$。e_i 到 e_j 的转移概率计算方法如下：

$$P_e(e_i,e_j)=\sum_{v\in V}\frac{h(v,e_i)}{\delta(e_i)}\frac{w(e_j)h(v,e_j)}{\sum_{\hat{e}\in E(v)} w(\hat{e})}=\frac{1}{\delta(e_i)}\sum_{v\in e_i\cap e_j}\frac{w(e_j)}{\sum_{\hat{e}\in E(v)} w(\hat{e})} \tag{5}$$

矩阵表示形式为：

$$P_e = D_e^{-1} H^T D_v^{-1} H W_e$$

D_e 为公式（3）中的超边的度对角矩阵，H^T 是关联矩阵 H 的转置矩阵，行为边，列为点。D_v 是公式（2）中的带权的度对角矩阵。W_e 为边权重对角矩阵，对角线上的值为对应边的权重。为了避免回路，我们把 P_e 中的对角线元素置为 0，然后再把 P_e 归一化，使每一行元素之和为 1。

我们采用 PageRank 方法来进行随机游走，$\vec{u}$ 为待排序的超边向量，α 为阻尼系数，具体如下所示：

$$\vec{u}_{(i+1)} = \alpha P_e^T \vec{u}_{(i)} + (1-\alpha)\vec{e}/n \qquad (6)$$

n 是超图的边的条数，$\vec{e} \in \mathbb{R}^{(n\times 1)}$ 是长度为 n 的单位向量。$\alpha P_e^T \vec{u}$ 表示漫游者从当前边 e 选择一个点跳转到另外一条边。$(1-\alpha)\vec{e}/n$ 表示漫游者以 $(1-\alpha)/n$ 的概率跳转到任意的其他边。

3.4 基于点的随机游走

基于点的随机游走过程与基于边的随机游走过程类似，漫游者的游走过程也分为两步：第一步，漫游者从与当前所在点 v_i 关联的所有边中以 $w(e)$ 与包含 v_i 的所有边的权重之和 $\sum_{\hat{e}\in E(v_i)} w(\hat{e})$ 的比值作为概率选择一条边 e；第二步，漫游者从 e 上随机选择一个点 v_j，且满足 $v_i, v_j \in e$。v_i 到 v_j 的转移概率计算方法如下：

$$P_v(v_i, v_j) = \sum_{e\in E} w(e) \frac{h(v_i,e)}{\sum_{\hat{e}\in E(v_i)} w(\hat{e})} \frac{h(v_j,e)}{\delta(e)} = \sum_{e\in E(v_i)\cap E(v_j)} \frac{1}{\delta(e)} \frac{w(e)}{\sum_{\hat{e}\in E(v_i)} w(\hat{e})} \qquad (7)$$

矩阵表示形式为：

$$P_v = D_v^{-1} H W_e D_e^{-1} H^T$$

D_e 为公式（3）中的超边度对角矩阵，H^T 是关联矩阵 H 的转置矩阵，行为边，列为点。D_v 是公式（2）中的有权的度对角矩阵。W_e 为边权重对角矩阵，对角线上的值为对应边的权重。为了避免回路，我们把 P_v 中的对角线元素置为 0，然后再把 P_v 归一化，使每一行元素之和为 1。

我们采用 PageRank 方法来进行随机游走，$\vec{v}$ 为待排序的点向量，α 为阻尼系数，具体如下所示：

$$\vec{v}_{(i+1)} = \alpha P_v^T \vec{v}_{(i)} + (1-\alpha)\vec{e}/m \qquad (8)$$

m 是超图的点的个数，$\vec{e} \in \mathbb{R}^{(m\times 1)}$ 是长度为 m 的单位向量。$\alpha P_v^T \vec{v}$ 表示漫游者从当前点 v 选择一条边跳转到另外一个点。$(1-\alpha)\vec{e}/m$ 表示漫游者以 $(1-\alpha)/m$ 的概率跳转到任意的其他点。

3.5 协同抽取方法

我们用两个列向量 $u = [u(s_i)]_{n\times 1}$ 和 $v = [v(t_i)]_{m\times 1}$ 分别表示一篇特定文章中句子和词的重要性得分。α 的值设为 0.85。u 和 v 的所有元素的初值分别设为 $1/n$ 和 $1/m$，即初始时

每个句子包含的主题信息密度为$1/n$，每个词包含的主题信息为$1/m$。我们迭代交替执行以下两步，直至u和v的值收敛。

1. 根据v的值，我们可以计算D_v，W_e，然后用下面的公式计算句子重要性得分：

$$P_e = D_e^{-1} H^T D_v^{-1} H W_e$$

$$u^{(n)} = \alpha P_e^T u^{(n-1)} + (1-\alpha)\vec{e}/n$$

2. 根据u的值，我们可以重新计算D_v，W_e，然后用下面的公式计算词重要性得分：

$$P_v = D_v^{-1} H W_e D_e^{-1} H^T$$

$$v^{(n)} = \alpha P_v^T v^{(n-1)} + (1-\alpha)\vec{e}/n$$

$u^{(n)}$和$v^{(n)}$分别表示句子得分向量和词得分向量在第 n 次迭代的计算结果。如果相邻两次迭代$u^{(n)}$和$u^{(n-1)}$，$v^{(n)}$和$v^{(n-1)}$对应位置的元素差值的绝对值小于某固定阈值（本文设为 0.0001），则停止迭代，同时我们得到了句子和词的最终得分，我们抽取得分最高的若干句子生成摘要，取得分最高的若干词作为文章的关键词。

4 实验与分析

4．1 摘要评价

4.1.1 数据集及评价标准

实验语料采用 NLPCC 2015 面向微博的新闻文本摘要任务数据集。该数据集包括 250 篇来自新浪的新闻文本，包括原始文本和已分句的文本，本实验采用后者。

评价方法采用 ROUGE[18]工具（1.5.5 版），此工具被 DUC 作为标准评测工具，在历年的自动文本摘要任务中被广泛使用。它通过计算待评价摘要与标准摘要在 n-gram 上的重叠度来衡量机器生成摘要的质量。ROUGE 可以分别计算 1、2、3、4-gram 和最大共现子序列 ROUGE-L 的分数。在这些分数中，基于 1-gram 的 ROUGE 分数（ROUGE-1）被公认为和人工评价的结果最接近[18]。在表 1 中展示了 ROUGE-1、ROUGE-2、ROUGE-3、ROUGE-4 和 ROUGE-L。另外由于摘要长度限定在 140 字以内，所以我们在评价时使用了“-l”命令。

4.1.2 评价结果

我们选取了四个 baseline 与本文提出的 HBCE 方法比较，这四个方法是：HBR (Hyperedge-based Ranking)、GBIR（Graph based Iterative Reinforcement）、SentenceRank 和 Log likelihood。HBR[19]是基于超边的排序方法，该方法首先使用 Topic Signatures[5]方法找出文章的主题词，每个主题词权重为 1，非主题词权重为 0，用公式（4）计算句子权重，然后用基于边的随机游走算法（即公式（5）（6））计算句子的最终得分。GBIR[3]方法是基于图的迭代强化方法，该方法在计算词与词之间的语义相似性用到的方法是基于语料的滑动窗口法，窗口大小设置为 5，把 250 篇新闻文本作为背景语料。SentenceRank[1]方法是 Mihalcea 和 Tarau 在 2004 年提出来的，该方法利用句子与句子之间的关系来对所有句子排序。Log Likelihood 方法是将 HBR 方法计算出带权重的句子，按权重大小排序，这里的句子权重为主题词密度。评价结果见表 1。

表 1 摘要评价结果

System	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
HBCE	**0.51470**	**0.33311**	**0.26062**	**0.22368**	**0.35144**
HBR	0.50542	0.32371	0.25219	0.21545	0.34892

SentenceRank	0.47839	0.28698	0.21580	0.18251	0.31949
GBIR	0.47673	0.28530	0.21361	0.17973	0.32591
Log likelihood	0.44941	0.25350	0.18186	0.14994	0.28135

4.1.3 结果分析

从实验结果来看，本文提出的 HBCE 方法在所有 ROUGE 指标上都优于其它方法，证明了此方法的有效性。HBR 方法在没有利用词的强化作用下，仅利用基于超边的随机游走算法，其结果比 SentenceRank 方法要好，证明句子与词之间的高阶关系是一个十分重要的关系，但是基于图的排序方法中无法表示这种关系。基于图的迭代强化方法 GBIR 和利用句子之间关系的图排序算法 SentenceRank 方法效果相当。Log Likelihood 按照句子主题词密度给句子排序，从表 1 可以看出这个方法是有效的，同时证明我们在 HBCE 方法中采用主题信息密度作为句子的权重是合理的。

4.2 关键词评价

4.2.1 评价方法

实验语料采用 NLPCC 2015 面向微博的新闻文本摘要任务数据集中的前 30 篇新闻文本，人工标注出每篇文本中的关键词，每篇最多标注 10 个关键词。对这 30 篇文本，利用本文提出的方法，每篇文本提取出 10 个得分最高的词。我们通过计算候选词在标注词上的正确率 P，召回率 R 以及 F 值（F=2PR/(P+R)）来评价方法的有效性。

4.2.2 评价结果

在实验中，分别用 GBIR、WordRank 和 Topic Signatures 为每一篇文档提取出 10 个候选词来与本文的方法作对比。WordRank[1]方法是 Mihalcea 和 Tarau 在 2004 年提出来的，该方法利用词与词之间的共现关系来对所有词排序。Topic Signatures[5]方法提取每篇新闻文本关键词的时候，把除去这篇文本的剩下 249 篇新闻文本作为背景语料。评价结果见表 2。

表 2 关键词评价结果

System	准确率	召回率	F 值
Topic Signatures	**0.483**	**0.524**	**0.501**
GBIR	0.473	0.510	0.490
HBCE	0.463	0.502	0.481
WordRank	0.440	0.477	0.456

4.2.3 结果分析

从表 2 中的对比实验结果可以看出，本文的方法比只考虑词与词之间关系的 WordRank 方法效果要好，但是和基于背景语料的方法相比还存在一定的差距，如果从计算的时间复杂度和是否使用背景语料来综合考虑，本文的方法优势较为明显。

5. 总结与展望

本文提出了一种基于超图的协同抽取方法 HBCE，可以同时对一篇文档生成摘要和抽取出关键词，该方法以句子作为超边，以词作为结点构建超图，在一个统一的超图模型下同时利用句子与词之间的高阶信息来生成摘要和关键词。最后的评价结果表明本文的方法在文本摘要上的效果要优于其他基于图的方法，同时优于仅考虑句子与句子间关系而忽视了词与句子间潜在的高阶关系的方法，关键词抽取的评价结果也优于仅考虑词与词之间关系的方法，虽然与基于语料或知识库的方法相比还存在一定的差距，但是本方法的综合性能优势较为明显。下一步我们打算把该方法应用于英文数据集，以验证本文方法的有效性。除此之外，本文在进行迭代的过程中，无论是基于点的随机游走过程还是基于边的随机游走过程，都是以

一个固定的概率在超边上选择点，下一步我们将验证以该词的权重占句子权重的比值作为概率来选择点的有效性。

参考文献：

[1] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]//Association for Computational Linguistics, 2004.

[2] Erkan G, Radev D R. LexPageRank: Prestige in Multi-Document Text Summarization[C]//EMNLP. 2004, 4: 365-371.

[3] Wan X, Yang J, Xiao J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction[C]//Annual Meeting-Association for Computational Linguistics. 2007, 45(1): 552.

[4] Hovy E, Lin C Y. Automated text summarization and the SUMMARIST system[C]//Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. Association for Computational Linguistics, 1998: 197-214.

[5] Lin C Y, Hovy E. The automated acquisition of topic signatures for text summarization[C]//Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000: 495-501.

[6] Nomoto T, Matsumoto Y. A new approach to unsupervised text summarization[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 26-34.

[7] Kupiec J, Pedersen J, Chen F. A trainable document summarizer[C]//Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1995: 68-73.

[8] Conroy J M, O'leary D P. Text summarization via hidden markov models[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 406-407.

[9] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998: 335-336.

[10] Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 19-25.

[11] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

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

[13] Turney P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2000, 2(4): 303-336.

[14] Wu Y B, Li Q, Bot R S, et al. Domain-specific keyphrase extraction[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 283-284.

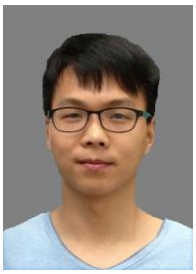
[15] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]//Proceedings of the fourth ACM conference on Digital libraries. ACM, 1999: 254-255.

[16] Mihalcea R, Csomai A. Wikify!: linking documents to encyclopedic knowledge[C]//Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007: 233-242.

[17] Grineva M, Grinev M, Lizorkin D. Extracting key terms from noisy and multitheme documents[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: 661-670.

[18] Lin C Y, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 71-78.

[19] Bellaachia A, Al-Dhelaan M. Multi-document Hyperedge-based Ranking for Text Summarization[C] //Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 1919-1922.



莫鹏（1990—），硕士研究生，主要研究领域为自然语言处理。
Email: mp@mails.ccnu.edu.cn

胡珀（1980—），博士，副教授，主要研究领域为自然语言处理。
Email: phu@mail.ccnu.edu.cn

黄湘冀（1965—），博士，教授，主要研究领域为信息检索。
Email : drjimmyhuang@gmail.com

何婷婷（1964—），博士，教授，主要研究领域为自然语言处理。
Email: tthe@mail.ccnu.edu.cn