DOI：10.7652/xjtuxb201512010

利用信息传播特性的中文网络新词发现方法

孙立远[1,2]，周亚东[3]，管晓宏[1,3]
（1. 清华大学智能与网络化系统研究中心，100084，北京；2. 国家计算机网络应急技术处理协调中心，100029，北京；3. 西安交通大学智能网络与网络安全教育部重点实验室，710049，西安）

摘要：针对已有方法识别出的网络中文新词生命周期短且很快不再为人们所用的问题，提出了一种基于信息传播特性的中文新词发现方法。该方法结合“新词传播范围广、持续时间长”的特点，从用户覆盖率、话题覆盖率和新词生命周期 3 个方面设计统计量；采用 *N*-gram 算法得到候选词串列表；用基于词频和词语灵活度的方法过滤垃圾词串。实验中以微博文本作为语料来源，与已有方法相比，用户特性使得新词识别的准确率提高了 11%，话题特性使准确率提高了 10%，时间特性使准确率提高了 13%，综合用户、话题和时间的方法使准确率提高了 16%。实验结果表明：该方法中的每个特性都提高了中文网络新词识别的准确率，而且同时考虑 3 种特性的准确率比只考虑单一特性的高。

关键词：新词发现；信息传播；用户行为；时间特性
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2015)12-0059-06

A Method of Discovering New Chinese Words from Internet Based on Information Propagation

SUN Liyuan[1,2], ZHOU Yadong[3], GUAN Xiaohong[1,3]
(1. Center for Intelligent and Networked Systems, Tsinghua University, Beijing 100084, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center, Beijing 100029, China;
3. MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A method of discovering new Chinese words from Internet based on information propagation is proposed to solve the problems that the recognizing results of existing methods always have short life cycles and will not be used again in soon. The method combines the characteristics of new words such as widely spreading and long lasting, and three statistics, i.e. coverage rate of users, coverage rate of topics and life cycle of a new word, are defined. The *N*-gram algorithm is applied to generate candidates of new words, then the word candidates are filtered bade on word frequency and word flexibility. Experiments with the text of microblogs as corpus and comparisons with the existing methods show that the user statistic enhances the accuracy rate of recognizing new words by 11%, the topic statistic enhances the accuracy rate by 10%, and the time statistic enhances the accuracy rate by 13%. When the three statistics are combined, the accuracy rate is raised by 16%. It can be concluded that each single statistic considered by the proposed method can enhance the accuracy rate, and more accurate rate can be obtained by considering the combination of the three statistics rather than just considering one statistic.

收稿日期：2015-07-10。 作者简介：孙立远(1986—)，女，博士生；周亚东(通信作者)，男，博士，讲师。 基金项目：国家自然科学基金资助项目(61221063,61572397,61502383)；陕西省自然科学基础研究计划资助项目(2015JM6298)。
网络出版时间：2015-09-21 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20150921.1442.006.html

Keywords: new word discovery; information propagation; user behavior; temporal characteristics

随着计算机网络技术的迅猛发展,网络日益成为社会信息发布和语言文化传播的平台,由此不断产生新的网络用语和热门词汇。一些认同度较高的网络新词逐渐被人们接受,并被扩充到汉语词汇中。由于散落在海量网络文本中的网络新词很难仅依靠人工进行查找、检索和统计,因此如何快速有效地自动检测网络数据并发现新词是一项亟需解决的问题。

目前,学术界对新词的定义尚不统一。有研究认为,只要是以前没有出现过的词就是新词[1-5]。但是,在这样的定义下识别出的新词绝大部分从出现到消失总的存在时间不超过 5 d,生命周期很短;只有 0.80%的新词生命周期达到 26 d 以上,能够被广泛使用[3]。因此,考虑到信息传播的特性,本文将存在时间久、使用范围广泛也作为判断新词的标准。

中文新词发现方法一般包括 2 个步骤:一是划分文本生成候选词串,这是因为中文文本中词与词之间没有明确的边界;二是从候选词串中发现新词。

对于第一步划分文本生成侯选词串,大多数方法采用概率词法分析系统(例如 ICTCLAS 等中文分词软件[2])或是基于词典查找的方法,但是这种方法创建和维护词典困难,而且因为新词并不在词典中,所以由基于词典的分词方法产生的候选词串不一定包括所有可能的新词,造成查全率的损失。

对于第二步从候选词串中发现新词,目前主要有两类方法:基于规则的方法和基于统计的方法[1]。基于规则的方法是指结合构词法、语义、词性等语言学特征创建匹配模板发现新词[6-7]。这类方法的优点是准确率高,但是规则维护困难,且适应性和移植性较差。基于统计的方法一般通过定义统计量将新词发现看做模式识别的二分类问题,其中两个类别分别代表"是新词"和"不是新词"。根据有无训练语料,基于统计的方法可以分为有监督方法和无监督方法。有监督方法首先提取语料中的词项特征,然后训练分类器判断是否是新词,这类方法依赖于训练语料和分类器设计;无监督的方法由于没有训练语料,因而采用设定统计量阈值的方法,如果一个候选词串满足统计量的阈值要求则被看作是新词。常用的统计量有词频[8]、互信息[9]、上下文信息熵[10]等。这类基于统计的方法,其优点是具有很强的适应性和可移植性,但是需要大量语料进行统计而且准确率相对较低。另外,已有的统计量并不能体现新词在传播范围和存在时间上的特点。

本文利用"新词传播范围广、持续时间长"的特点,提出基于信息传播特性的中文新词发现方法。该方法采用 N-gram 算法得到候选词串列表,用基于词频和词语灵活度的方法过滤垃圾词串,并结合信息传播特性从用户覆盖率、话题覆盖率和新词生命周期 3 个方面设计统计量。本文以近年来非常流行的网络微博应用为示例,采用微博文本作为语料来源。

1 基于信息传播特性的中文网络新词发现方法

1.1 N-gram 划分词串

因为微博文本的书写风格灵活,形成了一些特有的微博表达方式和使用方法,例如 URL、@符和表情符号等。这些微博表达方式中一般不包含新词,所以本文首先对微博语料内容进行自动预处理,过滤掉不包含新词的部分,以提高后续步骤的处理效率。其中,URL 短链接字符串和@用户名称采用正则表达式过滤,表情字符串采用表情符号列表过滤。

为了避免中文分词软件查全率不高的缺点,本文采用 N-gram 算法[11]划分预处理后的文本,顺次将临近的 N 个汉字聚集在一起形成一个候选词串。考虑到新词至少由 2 个汉字组成,而大于 5 个汉字的词语比例非常小,本文设定阈值 $N_{\max}$ 为 5。为了提高处理效率,N-gram 算法划分词串的同时统计每个词串出现的次数,具体的实现过程如下。

输入 预处理后的语料为 T,词串中的汉字个数为 N,初始值 $N_{\min}$,最大值为 $N_{\max}$

输出 候选词串集合

步骤 1 逐条读取语料 T 中的微博,按空格切割成 I 个孤岛词串。

步骤 2 扫描第 i 个孤岛词串,以连续 N 个字符的字符串 S 为候选,查找候选词串集合,如果 S 在候选集合中,则 S 的频次加 1;如果 S 不在候选集合中,则将 S 加入候选集合。

步骤 3 $i=i+1$,如果 i 大于 I,转至步骤 4,否则转至步骤 2。

步骤 4 $N=N+1$,如果 $N>N_{\max}$ 则退出,否则转至步骤 1。

N-gram 算法的优点是方法简单,容易实现,查全率高,能保证所有新词都在候选词串中,但缺点是产生大量无意义的垃圾词串。

1.2　基于词频和词语灵活度的过滤方法

本文采用基于词频和词语灵活度的方法过滤垃圾词串，以便提高后续基于统计的方法的效率。

一个可以被视为词的字符串，应该会被广泛使用，因此在语料中也会频繁出现。出现频率低的词串不大可能是有意义的词串。91.4%的候选词串出现次数小于等于 2，因此本文设定词频过滤方法的阈值为 2。

另外，根据中文的构词规则，有些字符不经常出现在词首或词尾，这些字符被称为停用字，包括词首停用字和词尾停用字。本文利用词语的灵活度（即每个字符构成词的概率）和设定的阈值比较，来发现停用字。用 c 表示待判断的字符，• 表示任意字符，c 可能出现在词首、词尾或是词中间，由此定义词首停用字为

$$R_{\text{first}}=\left\{c\mid c\in T,\frac{D(c\bullet)}{D(c\bullet)+D(\bullet c\bullet)+D(\bullet c)}\leqslant\alpha\right\} \tag{1}$$

定义词尾停用字为

$$R_{\text{last}}=\left\{c\mid c\in T,\frac{D(\bullet c)}{D(c\bullet)+D(\bullet c\bullet)+D(\bullet c)}\leqslant\alpha\right\} \tag{2}$$

式中：$D(S)$ 表示字符串 S 在语料中出现的次数；α 是字符出现在词首或词尾的概率阈值。实验中设定阈值为 0.1，共抽取出约 150 个停用字，包括“是”“的”“了”“们”等。

对候选词串过滤之后，本文采用统计的方法获得最终的新词结果。

1.3　基于信息传播特性的统计方法

本文新词发现的目标是使用范围广、存在时间长的未出现过的词。结合信息传播特性，有如下假设：如果使用某个词的用户数目越多，则说明该词的使用范围越广；如果某个词出现在越多的话题中，则说明该词的使用范围越广；如果某个词在一段时间内的频度变化是递增的，则说明该词更有可能长时间存在。所以我们从用户覆盖率、话题覆盖率和新词生命周期 3 个方面分别设计了用户特性统计量、话题特性统计量和时间特性统计量，最后综合这 3 个方面提出了综合统计量。

1.3.1　*用户特性统计量*　每条微博都有一个唯一的发布者，即微博用户，而一个用户可以发布不止一条微博。同一个用户的语言习惯固定，发表的微博内容在用词上也有相似性，但是绝大多数用户重复发帖的数目不多[12]，所以由于同一个用户语言使用习惯带来的偏差并不大。可以认为使用某个词的用户数目越多，该词的使用范围越广。

用三元组 (w,m,u) 表示候选词 w 出现在微博 m 中且微博 m 的发布者是用户 u，用二元组 (m,u) 表示微博 m 的发布者是用户 u，设计用户特性统计量为

$$U(w)=\frac{|\{u\mid m\in T,(w,m,u)\text{ 存在}\}|}{|\{u\mid m\in T,(m,u)\text{ 存在}\}|} \tag{3}$$

式中：分母表示所有微博语料中不同用户的数目；分子表示包含词 w 的微博中不同用户的数目。

1.3.2　*话题特性统计量*　如果某个词在很多类别的话题中出现，说明该词的使用范围很广。由于每个类别的话题数目不同，所以先对各类别的话题数归一化然后再统计某个候选词的话题特性统计量数值。

用三元组 (w,m,k) 表示候选词 w 出现在微博 m 中且属于话题 k。话题的类别用 K 表示，设计话题特性统计量为

$$T(w)=\sum_{K}(|\{k\mid k\in K,(w,m,k)\}|/D(K)) \tag{4}$$

式中：$|\{k|k\in K,(w,m,k)\}|$ 表示出现词串 w 的微博涉及的话题中属于类别 K 的个数；$D(K)$ 表示类别 K 中话题的总数。

1.3.3　*时间特性统计量*　候选词串如果存在时间越久越有可能是新词，如果候选词串的词频在增长则更有可能是新词。考察候选词串在一段时间内的频度变化趋势，通过评价函数给每个候选词串打分，并据此设计时间特性统计量。

候选词 w 的观测区间为语料中该词第一次出现的时间 $t_{w,\text{f}}$ 到该词最后一次出现的时间 $t_{w,\text{e}}$，则候选词 w 的观测天数为 $n_w=t_{w,\text{e}}-t_{w,\text{f}}+1$。第 i 天的词频为 $a_i(i=1,2,\cdots,n_w)$。定义评价函数为

$$f_w(a_{i+1},a_i)=\begin{cases}0, & \text{if } a_{i+1}=0\\ 0.5, & \text{if } a_{i+1}>0 \text{ and } a_{i+1}\leqslant a_i\\ 1, & \text{if } a_{i+1}>0 \text{ and } a_{i+1}>a_i\end{cases} \tag{5}$$

时间特性统计量定义为观测区间内评价函数数值的和

$$L(w)=\sum_{i=1}^{n_w-1}f_w(a_{i+1},a_i) \tag{6}$$

1.3.4　*综合统计量*　以上 3 种统计量从信息传播特性出发，各有侧重。为了提高新词发现的整体准确率，同时考虑这 3 种统计量，提出了综合统计量。

由于每个特性统计量的取值范围不同，不能直接相加，所以先对它们做归一化，使每个统计量的取值都在[0,1]之间。采用如下的归一化方法

$$Y_X = (X - X_{\min}) \big/ (X_{\max} - X_{\min}) \tag{7}$$

式中：$X \in \{U, T, L\}$ 分别代表用户特性统计量、话题特性统计量和时间特性统计量；$X_{\min}$ 表示变量 X 的最小取值；$X_{\max}$ 表示变量 X 的最大取值。

归一化之后，综合统计量为各个统计量的和

$$Q(w) = Y_U + Y_T + Y_L \tag{8}$$

2 实验与结果分析

2.1 数据介绍

利用新浪微博 API，随机选取新浪微博中粉丝数较多的账号作为采集起点，采用“滚雪球”策略，采集了 2013 年 3 月 1 日到 2013 年 5 月 31 日期间这些账号发布的每条微博的 ID 号、发布时间、发布人、内容等信息，经过去除垃圾微博等预处理后，构建了包括 68 754 名用户、107 万条微博的原始数据集。

通过识别每条微博中的话题标签生成研究中需要的话题数据集。在新浪微博中，用 2 个 # 标识一个话题，例如“#雅安地震#”和“#李宇春戛纳行#”等。本文首先识别微博中出现的所有话题标签，共 218 619 个，其中大部分话题包含的微博数很少，对新词识别的作用也有限，因此本文进一步识别传播范围较广的热门话题数据。考虑到新浪微博会公布每天的前 10 个话题，实验中选取 2013 年 3 月出现的微博数最多的 300 个话题标签，合并属于同一个话题的多个标签，生成话题列表。然后，在全部微博（其中包括未使用 # 标识但提及某个话题的微博）中逐个搜索话题列表中的话题，如果一条微博中出现多个话题标签，则标记第一个出现的话题标签作为这条微博所属的话题类别。另外，为了观测到话题的完整生命周期，实验中只保留了 2013 年 3 月 2 日以后出现的话题。最后，构建了包括 36 038 名用户、19.5 万条微博、涵盖 106 个最热门话题的话题数据集。参照新浪微博的话题类别划分方法，将 106 个热门话题分为 5 类，包括社会新闻类、广告公关类、网络热点类、电影电视类和其他类，见表 1。

表 1 话题类别概况

类别	话题数	用户数	微博数
社会新闻类	24	28 812	114 023
广告公关类	25	15 721	38 864
网络热点类	19	6 853	17 179
电影电视类	22	7 831	12 699
其他类	16	6 162	12 696

2.2 实验设置

下面介绍实验所用的评价指标和基准方法。

2.2.1 *评价指标* 因为微博数据量极大，很难标注出所有真实的新词，所以本文采用无监督的方法，提出基于信息传播特性的统计量，对每个候选词打分，分值越高则越有可能是新词，然后将每个候选词的分值从大到小排序，获得最终的新词列表。

本文方法和对比方法的识别正确率采用了信息检索领域常用的前 N 个结果的准确率（$P@N$）[3]来计算，具体来说就是对各个方法返回的前 N 个新词结果进行人工判别，判断‘是新词’或者‘不是新词’，把‘是新词’的比例作为前 N 个结果的准确率。N 一般取值 100，200，300 等（相应的表示为 $P@100$，$P@200$，$P@300$），以便减少人工标注的工作量。

2.2.2 *基准方法* 本文的基准方法包括：常见的经典统计量互信息量、邻接熵，以及最新的基于词内部结合度和边界自由度的方法。通过和基准方法的对比，说明各方法性能的优劣。另外，为了说明本文方法的有效性，基准方法使用的数据源和本文方法的数据源一致，并采用了相同的预处理过程。

基准 1 互信息量 MI 是衡量两个事件之间相关性的信息度量。对于候选新词 $w = c_1 c_2 \cdots c_n$，如果它的两个最长子串 $w_{\text{left}} = c_1 c_2 \cdots c_{n-1}$ 和 $w_{\text{right}} = c_2 c_3 \cdots c_n$ 的相关性越高，说明 w 越可能是一个词。本文使用文献[9]的计算方法，计算公式如下

$$B_{\text{MI}}(w) = \log(p(w) / p(w_{\text{left}}) p(w_{\text{right}})) \tag{9}$$

式中：$p(w)$ 是词 w 在所有候选词串中出现的概率。

基准 2 邻接熵 BE 利用信息熵来衡量候选新词 w 的左邻字符和右邻字符的不确定性[10]。不确定性越高，说明其上下文环境越丰富。用字符 x 和字符 y 分别表示 w 的左邻字符和右邻字符，则 w 的左邻熵 $H_{\text{left}}(w)$ 和右邻熵 $H_{\text{right}}(w)$ 的计算方法如下

$$H_{\text{left}}(w) = -\sum_x p(x \mid w) \log\, p(x \mid w) \tag{10}$$

$$H_{\text{right}}(w) = -\sum_y p(y \mid w) \log\, p(y \mid w) \tag{11}$$

邻接熵 $B_{\text{BE}}(w)$ 定义为左邻熵和右邻熵中较小的

$$B_{\text{BE}}(w) = \min\{H_{\text{left}}(w), H_{\text{right}}(w)\} \tag{12}$$

基准 3 词内部结合度和边界自由度 ICBF 由文献[13]提出，该方法对预处理后的语料进行中文分词，统计“散串”，并计算词内部的结合度（即互信息量），保留词内部结合度大于阈值的词语，最后计算词语

的左右边界自由度(即左邻熵和右邻熵)，把左右边界自由度都大于阈值的候选词作为新词。在本文的语料上，使用该方法共得到 1 982 个新词。为了更好地与本文方法比较并计算 $P@N$ 值，对于给定的 N，从这些新词中随机选择 N 个，然后判断新词的准确率，重复做 10 次取平均值作为该方法的 $P@N$ 值。

2.3　结果分析

实验中各种方法采用同样的新浪微博语料来发现中文新词。表 2 列出了本文方法和基准方法识别中文新词的准确率。对比结果显示，本文提出的每个特性都比基准方法的准确率高，而且同时考虑 3 种特性的准确率比只考虑单一特性的高。具体来说，和基准方法 1-MI 相比，用户特性使得前 100 个词的识别准确率提高了 11%，话题特性使准确率提高了 10%，时间特性使准确率提高了 13%，综合用户、话题和时间的方法使准确率提高了 16%；前 200 个词和前 300 个词的识别准确率也有类似的提高。由于微博中的词语使用较不规范，有大量昙花一现的词语，基准方法中只根据上下文的文本信息不能有效地甄别词语使用周期的差别，而且基准方法中 BE 比 MI 的准确率高，说明考虑上下文信息能提高新词识别的准确率。

表 2　新词识别准确率的对比

方法	准确率/%		
	$P@100$	$P@200$	$P@300$
基准 1-MI	20	11	7
基准 2-BE	35	16	10
基准 3-ICBF	28	13	9
方法 1-用户特性统计量	31	16	10
方法 2-话题特性统计量	30	15	10
方法 3-时间特性统计量	33	17	11
方法 4-综合统计量	36	18	12

另外，表 2 中准确率最高为 36%，并不是很高。这是因为微博词汇的随机性比较大，本文在处理过程中为了保证不漏掉可能的新词，不得不容忍了大量无意义词语的出现，因而影响了整体正确率。本文方法的用途是为发现新词提供数据输出，以减轻人工从大量文本中筛选的工作量，后期还可以通过人工的方式提高准确率。

表 3 列举了按统计量数值排名前 20 的结果中有意义的新词，这些词大多是近些年出现的，说明本文的中文网页新词自动获取方法能有效地识别出网络新词。另外，前 20 个结果中两字词的准确率最高，四字词的准确率最低。说明词串过滤步骤中，四字词的过滤效果最差，需要研究更有效的词串过滤方法。

表 3　新词发现结果举例

词语类型	新词发现结果
两字词	微博，转发，逆袭，给力，卧槽，城管，屌丝，腐女，抓狂，博主，亲们，靠谱，木有，牛逼，脑残，点赞，互粉，有爱
三字词	大尺度，光棍节，冷笑话，正能量
四字词	舞林争霸，网络公知
五字词	一次性筷子，地球一小时，白色情人节，习李氏理论，经济适用女

3　结　论

本文利用“新词传播范围广、持续时间长”的特点，提出基于信息传播特性的中文新词发现方法。该方法采用 N-gram 算法得到候选词串列表，用基于词频和词语灵活度的方法过滤垃圾词串，并结合信息传播特性从用户覆盖率、话题覆盖率和新词生命周期 3 个方面设计统计量，实现针对微博文本的新词发现方法，实验结果表明该方法提高了中文网络新词发现的准确率。

参考文献：

[1] 张海军，史树敏，朱朝勇，等. 中文新词识别技术综述 [J]. 计算机科学，2010，37(3)：6-10.
ZHANG Haijun，SHI Shumin，ZHU Zhaoyong，et al. Survey of Chinese new words identification [J]. Computer Science，2010，37(3)：6-10.

[2] 霍帅，张敏，刘奕群，等. 基于微博内容的新词发现方法 [J]. 模式识别与人工智能，2014，27(2)：141-145.
HUO Shuai，ZHANG Min，LIU Yiqun，et al. New word discovery in microblog content [J]. Pattern Recognition and Artificial Intelligence，2014，27(2)：141-145.

[3] 苏其龙. 微博新词发现研究 [D]. 哈尔滨：哈尔滨工业大学，2013.

[4] 杨辉. 汉语新词语发现及其词性标注方法研究 [D]. 上海：复旦大学，2008.

[5] 邹纲，刘洋，刘群，等. 面向 Internet 的中文新词语检测 [J]. 中文信息学报，2004，18(6)：1-9.
ZOU Gang，LIU Yang，LIU Qun，et al. Internet-oriented Chinese new words detection [J]. Journal of

Chinese Information Processing, 2004, 18(6): 1-9.

[6] SUI Zhifang, CHEN Yirong. The research on the automatic term extraction in the domain of information science and technology [C]// Proceedings of the 5th East Asia Forum of Terminology. Beijing, China: China National Institute of Standardization, 2002: 17-21.

[7] HIDEKI I. Japanese named entity recognition based on a simple rule generator and decision tree learning [C]// Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001: 314-321.

[8] 罗盛芬，孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究 [J]. 中文信息学报，2003，17(3)：9-14.
LUO Shengfen, SUN Maosong. Chinese word extraction based on the internal associative strength of character strings [J]. Journal of Chinese Information Processing, 2003, 17(3): 9-14.

[9] YE Yunming, WU Qingyao, LI Yan, et al. Unknown Chinese word extraction based on variety of overlapping strings [J]. Information Processing and Management, 2013, 49(2): 497-512.

[10] HUANG J H, POWERS D. Chinese word segmentation based on contextual entropy [C]// Proceedings of the 17th Asian Pacific Conference on Language, Information and Computation. Piscataway, NJ, USA: IEEE, 2003: 152-158.

[11] 孙立远，袁睿翕，卞小丁. 一种中文网页新词自动获取方法：中国，ZL 200910237979.3 [P]. 2011-06-01.

[12] 周亚东. 在线社会网络热点话题识别与动态传播建模与分析研究 [D]. 西安：西安交通大学，2011.

[13] 李文坤，张仰森，陈若愚. 基于词内部结合度和边界自由度的新词发现 [J]. 计算机应用研究，2015，32(8)：51-55.
LI Wenkun, ZHANG Yangsen, CHEN Ruoyu. New word detection based on inner combination degree and boundary freedom degree of word [J]. Application Research of Computers, 2015, 32(8): 51-55.

[本刊相关文献链接]

杨攀，桂小林，安健，等. 利用贝叶斯原理在隐私保护数据上进行分类的方法. 2015，49(4)：46-52. [doi：10.7652/xjtuxb201504008]

李刘强，桂小林，安健，等. 采用模糊层次聚类的社会网络重叠社区检测算法. 2015，49(2)：6-13. [doi：10.7652/xjtuxb201502002]

李长路，王劲林，郭志川，等. 两阶段密度意识子空间聚类模型. 2014，48(10)：108-114. [doi：10.7652/xjtuxb201410017]

李涛，肖南峰. 应用相似度测量的图离群点检测方法. 2014，48(8)：67-72. [doi：10.7652/xjtuxb201408012]

陈家旭，唐亚哲，胡成臣，等. 延迟容忍网络中基于地点偏好的社会感知多播路由协议设计. 2014，48(6)：13-18. [doi：10.7652/xjtuxb201406003]

张赛，徐恪，李海涛. 微博类社交网络中信息传播的测量与分析. 2013，47(2)：124-130. [doi：10.7652/xjtuxb201302021]

莫同，褚伟杰，李伟平，等. 采用超图的微博群落感知方法. 2012，46(11)：120-126. [doi：10.7652/xjtuxb201211022]

豆增发，高琳. 利用膜粒子群优化和信息熵的医学文本特征选择. 2012，46(4)：45-51. [doi：10.7652/xjtuxb201204008]

陈刚，蔡远利，穆静，等. 海量信息异常检测问题的异常概率排序算法. 2011，45(4)：36-40. [doi：10.7652/xjtuxb201104007]

刘京鑫，孙剑，孟德宇. 基于视觉原理的分类算法. 2010，44(10)：116-119. [doi：10.7652/xjtuxb201010022]

冯少荣，张东站. 高效的用户访问预测新算法. 2010，44(4)：28-33. [doi：10.7652/xjtuxb201004007]

李小虎，杜海峰，庄健，等. 基于小世界原理的模型降阶优化研究. 2009，43(1)：108-113. [doi：10.7652/xjtuxb200901024]

朱虎明，焦李成. 基于免疫记忆克隆的特征选择. 2008，42(6)：679-682. [doi：10.7652/xjtuxb200806007]

周亚东，孙钦东，管晓宏，等. 流量内容词语相关度的网络热点话题提取. 2007，41(10)：1142-1145. [doi：10.7652/xjtuxb200710004]

杜海峰，李树茁，Marcus W. Feldman，等. 基于先验知识与模块性的网络社区结构探测算法. 2007，41(6)：750-754. [doi：10.7652/xjtuxb200706026]

（编辑　武红江）