面向网页的主题概念挖掘

刘琼琼 左万利 王 英
(吉林大学计算机科学与技术学院 长春 130012)
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 网页主题挖掘对自然语言处理如网页文本分类、文摘自动生成、信息融合等具有重要意义。挖掘网页主题可以帮助用户更好地理解网页内容。尽管已有一些从普通文本中挖掘概念的工作,但其很少考虑单词所属标签和位置对单词权重的影响,且没有工作给出上述两种影响因子的计算方法。借助WordNet,将网页主题从词语扩展到概念层次,提出了使用词性标注和词义消歧确定网页中单词词义并充分利用标签影响因子和位置影响因子对网页正文文本特征进行权重修正的主题概念挖掘方法,给出了两种影响因子的计算公式。在DMOZ数据集上的实验结果表明,修正权重可以明显提高主题挖掘精度,最高可达到0.95。

关键词 词性标注,词义消歧,标签影响因子,位置影响因子,权重修正
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.013

Topic Concept Discovery for Web Pages

LIU Qiong-qiong ZUO Wan-li WANG Ying
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)
(Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

Abstract Topic discovery from Web page has an important impact on natural language processing, such as text classification, automatic abstract generation, information fusion etc. Mining Web page topics can help users better understand the content of Web pages. Although there are some papers discussing topic discovery from ordinary texts, few of them consider how the label a word belongs to and the location in which a word appears affect the weight of a word, and none of them gives calculation methods for the two impact factors. This article extended Web topics from words level to concepts level based on WordNet, used speech tagging to determine the POS of the words, used word sense disambiguation to determine the words' meaning in the pages, made full use of label impact factor and location impact factor to modify the weights of concepts, and proposed calculation formulas for calculating these two impact factors. Experimental results on DMOZ dataset show that, compared with un-adjusted weight method, the adjusted weights method can significantly improve topic mining accuracy, which can reach up to 0.95 in the best case.

Keywords Speech tagging, Word sense disambiguation, Label impact factor, Location impact factor, Adjusted weights

文本主题挖掘是指从给定的文本中挖掘出其描述的关键信息,是数据挖掘的一个重要分支。主题挖掘被广泛应用于文本分类、文摘自动生成、信息融合等领域。

现有的主题挖掘技术包含以下几个方面:文献[1]提出了语义相似度和文本聚类相结合的主题发现方法;文献[2]从信息检索的角度,将查询词提交到搜索引擎,对搜索引擎返回的网页通过挖掘频繁项集的方法进行子主题挖掘,产生与查询主题相关度较高的网页;文献[3]基于LDA模型进行主题词挖掘,并结合背景语料扩充主题词;文献[4]提出了一种基于跨文档字频模式的几何形状主题建模算法,该算法依赖于数据和随机预测;文献[5]基于文本内容,挖掘论坛中具有高影响力的主题,通过对每一个单词赋予合理的权重,挖掘出高频词和关键词;文献[6]提出了含有语义信息的STU-DOM树模型,通过对HTML文档转化的STU-DOM树进行过滤和剪枝提取网页主题信息;文献[7]提出了使用Twitter列表发现主题的方法;文献[8]通过执行LDA然后进行聚类的方法来进行非监督的主题提取。现有的主题挖掘算法对于网页HTML代码标签和单词在网页中出现的位置没有给出定量的衡量方法,且大多局限于词语层面。因此,将主题挖掘扩展到概念层面,并综合考虑单词所属的HTML代码标签和单词在网页正文文本中的位置进行概念权重的定量分析是必要的。

到稿日期:2014-02-12 返修日期:2014-04-15 本文受国家自然科学青年基金项目(20130206051GX),吉林省重点科技攻关项目(20130206051GX)资助。

刘琼琼(1991—),女,硕士,主要研究方向为数据挖掘,E-mail:liuqq12@mails.jlu.edu.cn;左万利(1957—),男,教授,博士生导师,主要研究方向为自然语言处理、Web数据挖掘,E-mail:wanli@jlu.edu.cn(通信作者);王 英(1981—),女,博士,讲师,主要研究方向为本体、Web挖掘,E-mail:wangying2010@jlu.edu.cn。

本文提出了一种从给定待测网页出发提取网页主题概念的方法。该方法首先提取网页正文，结合 WordNet[1] 对网页正文文本进行词义消歧，确定其中所有名词的词义，即概念，使用其同义词集对相应概念进行扩展，统计扩展后的词集中所有单词出现次数之和作为该概念的初始权重，结合词语所属 HTML 标签和词语在网页正文文本中出现的位置进行概念权重修正，获得网页最终文本的概念级向量表示。选取权重最大的前 n 个概念作为网页最终主题概念集。本文算法流程如图 1 所示。

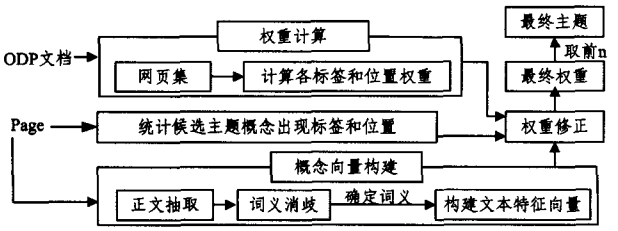

图 1　主题概念挖掘流程

1　特征向量构建

仅使用单词构造的文本特征向量缺乏必要的语义信息。鉴于此，本文采用词语同义词集构建初始文本特征向量。使用斯坦福大学的词性标注工具[2] 对网页正文中的词语进行词性标注，对标注后的词语使用文献[9]中提到的 Extended Gloss Overlops 方法，参考文献[10-12]借助 WordNet 对预处理后的网页文本进行词义消歧。在关于 29 个名词的 1754 个实例中与其他 13 种方法相比，该方法达到了最高的 F-measure值。它通过计算两个概念的 gloss 中重叠短语的单词数的平方和得到任意两个概念的相关度。

对一个目标单词词义的确定，首先需要计算目标单词的每一个词义与消歧窗口中其他词语的相关度，然后选取相关度总和最大的词义作为目标单词在该网页中的词义。其中消歧窗口是指词义消歧的范围，例如窗口可以是一个句子，那么词义消歧就是通过计算句中每一个单词与句子中其他单词的相关度来确定单词词义。目标单词 *word* 的词义 *WSense* 由式(1)确定。

$$WSense(word)=\arg\max_{i=1}^{m_t}\sum_{j=1,j\neq t}^{n}\max_{k=1}^{m_j}Relation(C_{ti},C_{jk}) \tag{1}$$

式中，m_t 为目标单词 *word* 的词义个数，t 为 *word* 在消歧窗口中的编号，m_j 为消歧窗口中第 j 个单词的词义数，n 为消歧窗口的大小，C_{ti} 和 C_{jk} 分别为第 t 个单词的第 i 个词义与第 j 个单词的第 k 个词义，$Relation(C_{ti},C_{jk})$ 为概念 C_{ti} 和 C_{jk} 的相关度。

主题词大多都是名词或者名词性短语，因此本文选取网页预处理后正文中所有名词作为候选主题词。由式(1)确定所有候选主题词的词义，通过 WordNet，用单词的同义词集中的词语扩展各候选主题词。用同义词集表示给定待测网页 p 的初始文本特征向量 W，如式(2)所示。

$$W=((S_1,w_1),(S_2,w_2),\cdots,(S_m,w_m)) \tag{2}$$

式中，m 为特征项个数，S_i 为 p 中第 i 个特征项经词义消歧后的同义词集，w_i 为 S_i 中各单词出现频度之和。

另外，网页正文中可能会出现一些非登录词，对于这些在 WordNet 中不存在的词，本文不予处理。对一篇网页进行词义消歧确定词义的算法描述如表 1 所列。

表 1　网页词义消歧算法描述

```
输入：网页 HTML 文档
输出：网页中名词的词义 sense[]
1. 提取网页正文为 context；
2. 使用"?"、"!"和"."对 context 进行分句得到数组 sentence[]；
3. While sentence 存在未处理项 do
4.    将 sentence[i]划分为单词数组 words[]和词性数组 pos[]；
5.    While words 存在未处理项
6.       IF pos[j]是名词且 words[j]在 WordNet 中 Then
7.          计算 words[j]的每一个词义与 sentence[i]中其他名词的相关度；
8.          依据式(1)得到 words[j]的词义 sense[j]；
9.       END IF
10. END
```

2　权重修正

网页文本主要由 HTML 代码组成，为半结构化数据，它不仅含有普通文本具有的文字信息，同时还包括文字链接、图片链接等超链接，还可以借助字号大小、粗体、斜体和下划线等标签突出关键词。另外，网页文本也同时具有与普通文本相似的特征，如主题词都倾向于出现在文本标题、正文开头和结尾等位置。因此，依照普通文本特征向量构建方法构建的特征向量并不全面，本文通过结合关键词所属 HTML 标签以及它在网页正文文本中出现的位置对初始权重向量 W 进行修正，以获得更精确的文本特征向量表示。

2.1　标签影响因子

网页一般都包含网页标题，该标题存在于"〈title〉"标签中，而网页中也常使用"〈h1〉、〈h2〉、〈b〉、〈strong〉、〈i〉、〈u〉"等标签突出关键词[3]。不同标签的重要度不尽相同，如〈title〉标签的重要度要高于大部分的其他标签，这是因为为了提高被检索的概率，网页标题中的词语通常是文章所要表达的主要信息。本文基于 DMOZ 分类目录，使用 DMOZ 分类自身所有的类别链中的各词作为网页主题词，采用统计手段，对若干网页中标签出现情况进行统计，获得网页集中出现标签 l 的网页总数 N_l、网页集大小 $totalN$、标签 l 中包含主题词总次数 c_l、标签 l 在网页集中出现的总次数 n_l，那么标签 l 在网页集中的影响因子由式(3)计算得到。

$$W_l=1+\frac{N_l}{totalN}\times\frac{c_l}{n_l} \tag{3}$$

即当出现统计数据中未曾出现的标签时，其重要度为 1；否则，其重要度大于 1。考虑标签在网页集中的出现比例($N_l/totalN$)可以减少个别标签仅出现在个别网页中而导致标签重要度过高的特殊情况的干扰。而 c_l/n_l 使用标签 l 中包含主题词总次数占所有主题词出现次数的比例，可以体现主题

[1] http://wordnet.princeton.edu/

[2] http://nlp.stanford.edu/software/tagger.shtml

词出现在标签 l 中的概率，从而体现标签 l 对主题词的影响力。

另外，网页 HTML 文本中有些主题信息作为标签属性，不在网页中显示，但这些信息对表现主题也可能具有重要作用，如 meta 标签中的 content 属性值，该属性值中常包含网页关键字、简介等信息便于搜索引擎查找信息和分类，我们将这类标签称为不可见标签；而标签中的内容可以被看见的标签称为可见标签。若仅对网页正文文本进行处理，那么不可见标签中包含的信息就会被忽略，这是不全面的。本文在对文本特征向量进行权重修正时，除考虑可见标签的信息外，也计算不可见标签信息。不可见标签中候选主题词的词义由式(4)确定。

$$Sense(word)=\arg\max_{i=1}^{m} T(word,i) \tag{4}$$

式中，$T(word,i)$ 为单词 $word$ 的第 i 个概念经可见标签权重修正后的权重，m 为单词 $word$ 的名词词义数。

Step1　对于当前网页中的词语，匹配词语之前第一个非成对标签的开始标签作为该词语所属标签。

例如有部分 HTML 网页代码“〈p class＝" more"〉〈strong〉 Disclaimer:〈/strong〉 Specifications …〈/p〉”，对于单词“specifications”，左侧第一个标签为“〈/strong〉”，该标签为结束标签；左侧第二个标签为“〈strong〉”，该标签在单词“specifications”前有结束标签，为成对标签；左侧第三个标签为“〈p〉”，该标签为单词“specifications”之前第一个非成对的开始标签。以上 3 个标签只有“〈p〉”满足条件，故单词“specifications”所属标签为“〈p〉”。

Step2　统计当前网页中包含主题词的标签出现次数、各标签中出现主题词的次数以及各标签出现总次数。

Step3　将当前网页的统计信息加入到目前为止所有已统计网页对应的信息中。

Step4　重复 Step1－Step3，直到所有网页信息均统计完成。依据式(3)计算网页集中所有标签的权重。

2.2　位置影响因子

不同位置上的词语是主题词的概率也不尽相同，如多数文章都会在开头对该文的主要内容进行描述，在结尾进行总结，因此，文章开头和结尾出现主题词的概率可能高于其他位置。本文基于 DMOZ 分类目录，抽取若干主要内容为文字的网页，使用 DMOZ 分类自身所有的类别链中的各词作为网页主题词集合，统计集合中各单词在正文 L_i 处出现的概率，获得位置影响因子 $LW(L_i)$，如式(5)所示。

$$LW(L_i)=\frac{N_{L_i}}{N_{total}} \tag{5}$$

式中，N_{L_i} 为网页集中主题词在 L_i 段的出现次数，N_{total} 为网页集中主题词出现总次数。使用各位置主题词出现次数占总次数的比例，可以体现该位置中词语为主题词的概率，也即该位置对主题词的影响力。

词语 i 在网页正文文本中的位置 L_i 的定义如式(6)所示。

$$L_i=\left\lceil\frac{D_i\times n}{Len}\right\rceil \tag{6}$$

式中，D_i 为词语 i 距正文开头的距离(以单词数计)，Len 为正文总单词数，n 为网页正文分段的段数。

主题词在网页中的出现位置权重获取算法如表 2 所列。

表 2　位置权重获取算法描述

输入：DMOZ 文档，网页正文分段数 n
输出：各分段主题词出现总次数 totalNum[]，网页集中所有主题词出现总次数 total

```
1. 选择 DMOZ 文档中主题包含 business 的 1000 个正文单词数大于 50 的网页 pages[];
2. While pages 中存在未处理网页 do
3.   获得 pages[i]主题词链 topics[];
4.   对 pages[i]提取正文 context;
5.   将 context 按照单词分为 n 段 segment[];
6.   For i←0 to n do
7.     While topics 未遍历完 do
8.       IF segment[j]包含主题词 topics[j] Then
9.                 totalNum[i]++;
10.      END IF
11.   END For
12. END While
13. total←sum(segment[0…n−1]);
14. END
```

2.3　权重修正

由 2.1 节中的分析可知，网页正文文本中概念不仅出现在正文可见标签中，还可能出现在不可见标签中，因此，对于标签影响因子对概念权重的影响，需要综合考虑可见标签和不可见标签两种情况。而依据经验，主题通常在文章开头和结尾出现次数较多。因此，结合标签权重和位置权重，概念 con 的最终权重如式(7)所示。

$$W_{final}(con)=\sum_{j=1}^{n} NLoc_j(con)\times LW(j)+W_{lable}(con) \tag{7}$$

式中，n 为正文分段的段数，$NLoc_j$ 表示在第 j 段 con 的出现次数，$LW(j)$ 表示第 j 段的权重，$W_{lable}(con)$ 为概念 con 在标签影响因子的作用下修正后的权重。$W_{lable}(con)$ 由式(8)计算得到。

$$W_{lable}(con)=\sum_{i=1}^{k} NLab_i(con)\times W_i(con)+\sum_{j=1}^{m} NLab_j(con)\times W_j(con) \tag{8}$$

式中，k 为出现概念 con 的可见标签个数，$NLab_i$ 表示可见标签 i 中 con 的出现次数，$W_i(con)$ 表示标签 k 的权重，m 为出现概念 con 的不可见标签个数，$NLab_j$ 表示可见标签 j 中 con 的出现次数，$W_j(con)$ 表示标签 j 的权重。

该权重修正公式将标签影响因子和位置影响因子对概念权重的影响结合在一起，充分考虑了概念所属标签和概念所在位置对概念权重的影响；同时将标签影响因子区分为可见标签的影响和不可见标签的影响，充分利用了不同类型标签对概念权重的不同影响，使得最终权重更为精确。

3　确定主题概念

鉴于停用词基本不存在于 WordNet 中，处理网页正文时不需要进行去除停用词的操作。因此，对于给定的网页 HTML 文档，首先要对该文档进行正文抽取，获取网页正文文本，然后对网页正文进行词义消歧处理，构建初始文本概念向量，然后依据标签影响因子和位置影响因子使用式(7)对初始文本向量进行权重修正，得到文本最终向量表示。依据上述步骤对待测网页进行主题概念挖掘的算法描述如表 3 所列。

表 3　主题概念挖掘算法描述

输入：网页 HTML 文档 page，常数值 n，标签权重列表 lable，位置权重列表 locate，段数 m

输出：主题概念集

```
1. sense[]←WSD(page);
2. 新建 list 列表用于存放每一个概念及其在网页中的总权重;
3. While sense 存在 list 中没有的项 do
4.    查找 sense[i]在 page 正文中所属标签 lab 和 lable 中该标签权重 lableWeight1;
5.    IF sense[i]不在 list 中 Then
6.       将 sense[i]加入 list，当前权重为 lableWeight1;
7.    ELSE
8.       sense[i]权重加 lableWeight1;
9. END While
10. 依据式(4)使用不可见标签权重修正 list 中个别项;
11. While list 存在未处理项 do
12.    使用段数 m 将网页正文划分为 m 段，判断 list[i]出现的段数，依据 locate 获得该段权重 locateWeight;
13.    list[i]权重加 locateWeight;
14. END While
15. list←sort(list);
16. 取 list 的前 n 项作为网页主题概念集;
17. END
```

4　实验

为了验证本文算法的正确率和权重修正的必要性，选取 DMOZ 开放目录下 business 的 5 个类别中若干网页进行主题概念挖掘，使用“挖掘正确的网页数/该类别网页总数”作为评价指标。下面从标签权重和位置权重的计算、网页中各候选概念的权重修正和最终主题概念选择几个方面给出实验结果。由于网页不同于普通文本，它的主题词可能以链接、图片等形式呈现，且噪音信息通常不包含主题词，因此，除词义消歧和位置权重计算使用网页正文文本外，其余部分选取网页 HTML 文本中的所有信息进行处理。

4.1　标签权重计算

本文选取 DMOZ 分类目录中 92 个关于 business 的主题共 1000 篇文本单词总数不小于 50 的网页进行统计，权重最大的前 12 个标签的出现情况如表 4 所列。

表 4　部分标签主题词统计概况

标签名称	主题词出现次数	标签出现总次数	比率	出现比例	权重
meta	4925	5137	0.9587	0.965	1.9251
title	823	1057	0.7786	0.993	1.7732
script	3488	10013	0.3483	0.891	1.3104
p	4143	15699	0.2639	0.926	1.2444
img	4028	15975	0.2521	0.957	1.2413
a	12973	57768	0.2245	0.991	1.2225
iframe	306	294	1.0408	0.130	1.1353
br	1905	15882	0.1199	0.857	1.1028
div	3761	40165	0.0936	0.918	1.0860
span	3146	30498	0.1032	0.772	1.0748
link	372	5007	0.0743	0.867	1.0644
h1	113	917	0.1232	0.424	1.0522

从表 4 可以发现以下几点特征：1）meta、title 标签的权重要明显高于其他标签，而 meta 标签权重最高，这验证了计算主题概念在不可见标签中出现权重的必要性；2）img、a、link 标签权重也较高，说明主题词倾向于以图片和链接的形式出现；3）iframe 标签的比率大于 1，这是由于该标签出现一次，而其中包含多个主题词的情况导致的，通过与出现比例相乘，避免了该标签因出现在个别网页中而导致权重过大的问题。

另外，待测网页文本中的某些标签可能并未在统计结果中出现过，对于这些标签，权重为 1；否则，根据标签权重计算公式，其权重大于 1。

4.2　位置权重计算

通过选取 DMOZ 分类目录中 92 个关于 business 的主题共 1000 篇文本单词总数不少于 50 的网页进行统计，将网页正文分为 50 段，得到如图 2 所示的统计结果。

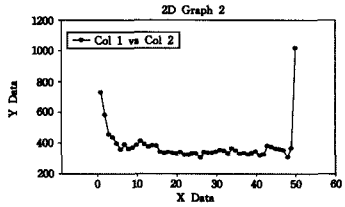

图 2　位置信息统计概况

从图 2 可以看出，主题词在网页正文文本的开头和结尾出现次数最多，明显多于中间部分，而在网页正文文本的中间部分主题词出现次数基本相等。由此可见，在文章开头和结尾出现的概念权重大于其他部分。

4.3　实验结果及分析

网页 HTML 文本区分为文本内容和标签内容，分别统计可见标签的文本内容中候选主题概念出现情况和不可见标签的标签属性中候选主题概念出现情况。结合标签权重和位置权重得到所有候选主题概念的最终权重，选择经排序后权重最大的前 n 个主题概念作为最终主题。

本文选取 DMOZ 分类目录中 business 主题的 5 个子类中的部分文档进行实验，选取的实验测试集的具体情况见表 5，其中主题概念链中单词词义由主题词链内词义消歧确定。

表 5　实验测试集概况

类别名称	文章总数	主题概念链
aluminum	20	Business#7 Materials#1 Metals#1 Aluminum#0 Consumer#0 Goods#0 Services#14 Watercraft#1 Boats#0
coffee	19	Business#0 Food#0 Products#1 Beverages#0 Coffee#2 Hospitality#0 Restaurant#0 Chains#2
exchange	25	Business#1 Transportation#4 Logistics#0 Marketplaces#0 Freight#1 Exchanges#3 Trucking#0 Stocks#3 Bonds#1
interest	20	Business#1 Publishing#0 Printing#1 Books#0 General#0 Interest#0
shipping	18	Business#7 Transportation#4 Logistics#0 Services#8 Auto #0 Transport#2 Motorcycle#0 Shipping#0 Goods#2 Packaging#1 Cases#17 Containers#0

实验结果是否正确通过如下方式确定：定义 DMOZ 中网页所属主题概念链的形式为$\langle c_1, c_2, \cdots, c_t \rangle$，其中 c_j 为类别名称，c_{j+1}是 c_j 的子类别。使用 WordNet 对 $c_1, c_2, \cdots, c_t$ 进行消歧，确定词义，并扩充 c_j 为其扩展词集 B_j（其中 B_j 为 c_j 的同义词集构成），则类别信息 C 扩展为：$B=\langle B_1, B_2, \cdots, B_t \rangle$。若待测网页的最终主题词的同义词集包含 B 中的任意一项，则主题概念挖掘正确，否则错误。

本文通过在表 5 所列的网页中对不进行权重修正的主题挖掘算法（NoAdjust）和进行权重修正的主题挖掘算法（Adjust）两种情况下各分类的正确率进行对比，表明了权重修正的必要性，比较示意图见图 3、图 4。其中图 3 为选取排序后权重最大的前 3 和前 5 个概念作为主题概念时，Adjust 和 NoAdjust 两种情况下准确率的对比；图 4 为选择前 3、前 5 和前 10 个概念作为主题概念时，本文算法（Adjust）的准确率对比。

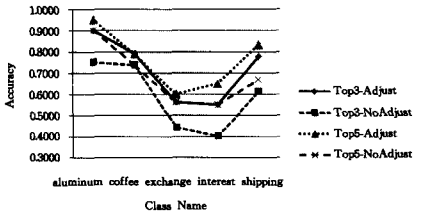

图 3　Top3、Top5 主题挖掘比较示意图

从图 3 可以看出，选择前 3 个概念或前 5 个概念作为最终主题的两种情况下使用标签和位置对特征向量进行权重修正(Adjust)比不使用标签和位置进行权重修正(NoAdjust)的准确率高；且同样进行权重修正时，其准确率差异在主题概念选择较少时表现更为突出。

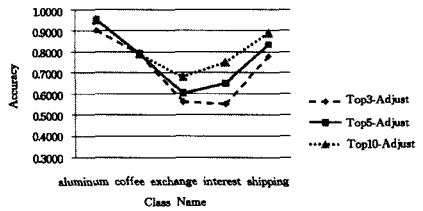

图 4　Top3、Top5 和 Top10 主题挖掘比较示意图

从图 4 可以看出，随着最终所选择主题概念数的增加，主题挖掘正确率增加，且在选择前 10 个概念作为主题概念的情况下，各类别准确率都在 0.68 以上，最高达到 0.95。

对于 n 值的选择，通过实验可以看到：随着 n 值的增大，主题概念挖掘的准确率也提高。但是选择的 n 个概念中非主题概念的个数也随之提升，且能表示网页主题的概念数一般不会太多，所以 n 值应该在一个较小且包含主题概念较多的范围内。从图 4 可以看出，在 n 取 5 时，算法准确率已经较高，因此，最终的 n 值可以取 5，并依据具体数据进行调整，以提高主题挖掘的准确率，并减少噪声出现的次数。

结束语　本文提出了以概念代替单词构建文本特征向量并使用标签和位置对待测网页中的候选主题概念进行权重修正得到最终主题的方法，分别给出了标签和位置权重计算公式，将标签分为可见标签和不可见标签两种情况，充分利用了网页的 HTML 代码蕴含的信息。通过实验验证了根据标签和位置修正网页文本特征的必要性，且主题挖掘具有较高的准确率。对于中文网页，可以使用中科院提供的分词和词性标注工具对中文网页正文进行分词和词性标注，并基于 Hownet 进行词义消歧。但受到词义消歧效率的限制，本文进行实验的数据集较小，尚需改进。

参 考 文 献

[1] Jayabharathy J, Kanmani S, Parveen A A. Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature[C]// 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN). 2011: 425-429

[2] Uluhan E, Badur B. Development of a Framework for Sub-Topic Discovery from the Web[C]// PICMET 2008 Proceedings. July 2008: 878-888

[3] Shi Jing, Li Wan-long. Topic Discovery Based on LDA Model with Fast Gibbs Samping[C]// 2009 International Conference on Artificial Intelligence and Computational Intelligence. 2009: 91-95

[4] Ding W, Rohban M H, Ishwar P, et al. Topic Discovery through Data Dependent and Random Projections[C]// International Conference on Machine Learning (ICML'13). 2013: 471-479

[5] Yang Yun, Wu Ya-nan. Content-based topic discovery of high-impact model[C]// 2010 2nd International Conference on Computer Engineering and Technology. 2010

[6] 王琦，唐世渭，杨冬青，等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展，2004，41(10)：1756-1792

[7] Yamaguchi Y, Amagasa T, Kitagawa H. Tag-based User Topic Discovery using Twitter Lists[C]// 2011 International Conference on Advances in Social Networks Analysis and Mining. 2011: 13-20

[8] Cheng L. Unsupervised topic discovery by anomaly detection [D]. Monterey, California: Naval Postgraduate School, 2013

[9] Pedersen T, Banerjee S, Patwardhan S. Maximizing semantic relatedness to perform word sense disambiguation[J/OL]. http://www.patwardhans.net/papers/pedersenBP05.pdf

[10] Naskar S K, Bandyopadhyay S. Word sense disambiguation using extended wordnet[C]// Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07). 2007: 446-450

[11] Naskar S K, Bandyopadhyay S. JU-SKNSB: extended WordNet based WSD on the English all-words task at SemEval-1[C]// Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics. 2007: 203-206

[12] Shen Wan, Angryk R A. Measuring semantic similarity using wordnet-based context vectors[C]// IEEE International Conference on Systems, Man and Cybernetics, 2007(ISIC). 2007: 908-913

(上接第 46 页)

[4] Zhang J, Ackerman M, Adamic L. Expertise networks in online communities: structure and algorithms[C]// Proc. of the 16th Int. Conf. on World Wide Web (2007). 2007: 221-230

[5] Tang J, Sun J, Wang C, et al. Social Influence Analysis in Large-scale Networks[C]// Proc. of the 15th Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD 2009). Paris, France, 2009: 807-816

[6] 田军伟. 基于社会网络的用户兴趣模型研究[D]. 成都：电子科技大学，2010

[7] Sharifi B, Hutton M-A, Kalita J K. Experiments in Microblog Summarization[C]// IEEE Second International Conference on Social Computing. 2010: 49-56

[8] Yang Y. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval[C]// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1994). 1994: 13-22

[9] Cortes C, Vapnik V. Support vector networks [J]. Machine Learning, 1995, 20: 273-297

[10] Feng Hao-di, Chen Kang, Deng Xiao-tie, et al. Accessor Variety Criteria for Chinese Word Extraction[J]. Computational Linguistics, 2004, 30(1): 75-93

[11] 沈崇玮. 基于微博数据的用户影响力分析研究[D]. 北京：北京邮电大学，2013