doi:10.6043/j.issn.0438-0479.2015.02.019

基于同义词词林和《知网》的短语主题提取

曾 聪,张东站*
(厦门大学信息科学与技术学院,福建 厦门 361005)

摘要:提出了利用主题词存在与否的基于主题词的短语抽取算法,并在其基础上利用社会知识词簇集合作为分类信息,词的相似度作为距离权重,利用改进 K 最近邻分类算法(KNN)的分类思想,提出基于《知网》词相似度的短语主题抽取算法.并在其基础上提出一种根据中文表达习惯的基于加权主题词的短语主题抽取算法.实验结果表明,后两种算法对短语主题抽取效果良好,平均查全率分别达到 78.88%和 83.39%,平均查准率达 99.06%和 99.70%.

关键词:短语主题;主题抽取;同义词词林;知网
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2015)02-0263-07

主题抽取是文本自动处理的基础工作之一,主题抽取通常针对的对象是一篇完整的文章,文章中包含着词、句、段等对文本主题有着不同贡献的信息.而主题抽取则是利用这些信息,对中文文章进行主题抽取.抽取步骤通常应用各种加权算法,有的基于词的绝对频率[1]、相对频率[2]、文中出现的位置[3]进行加权计算,还有的根据文章与文章之间的引用关系[4]进行计算,近年来,国内的学者也对文本主题抽取进行研究[5-8].

短语是搜索引擎的主要输入信息,研究短语的主题可以更好的对用户的搜索意图进行判断,将短语抽象出主题可以将用户输入的查询表层字符信息上升到主题层面,采取不同的主题形式来形式化地表示查询背后的搜索意图,从多个角度理解查询意图,并基于查询意图提供新颖的搜索服务与搜索模式[9].现在的主题抽取算法大都是基于统计和经验的加权体系.由于短语包含的信息与一篇完整的文章相比有着巨大的不同,所以基于统计和经验的加权体系无法直接应用于短语的主题抽取.

短语的主题往往包括在短语的词汇中,它或者是一个主题词或者是某一个主题词的同义词.利用同义词构造分类信息,短语的主题抽取可以转换成短语主题分类.

基于以上思想,本文提出了利用主题词存在与否的基于主题词的短语抽取算法,并在其基础上利用社会知识词簇集合作为分类信息,词的相似度作为距离权重,利用改进 K 最近邻分类算法(KNN)的分类思想提出基于《知网》词相似度的短语主题抽取算法.并在其基础上提出一种根据中文表达习惯的基于加权主题词的短语主题抽取算法.实验结果表明,后两种算法对短语主题抽取效果良好.

1 基于主题词的短语主题抽取算法(KWPSE)

短语由词汇构成,一个短语可以解释为词汇集合.短语的主题包含于词汇集合中,词语表达形式的多样化导致相同主题的短语所表现的主题词不一样.因此,构造候选主题词集就成为了短语主题抽取的第一部分.

1.1 候选主题词集确定

定义 1 定义 W 表示词的集合,w 表示一个词,即 $w\in W$.

定义 2 短语 P.短语可以看成是词语的集合,短语为词的集合的子集,即短语 $P\subset W$.

定义 3 词的主题 wt.它表示人们对某个词的一种直观认识.这种认识也是一个词,即:$wt\in W$.一个词 w 可能有多个主题.

收稿日期:2014-04-29 **录用日期**:2014-08-25
基金项目:国家自然科学基金(61303004);福建省自然科学基金(2013J05099)
* **通信作者**:zdz@xmu.edu.cn
引文格式:曾聪,张东站.基于同义词词林和《知网》的短语主题提取[J].厦门大学学报:自然科学版,2015,54(2):263-269.
Citation:Zeng Cong, Zhang Dongzhan. Phrase subject extraction based on synonyms and HowNet[J]. Journal of Xiamen University: Natural Science, 2015, 54(2): 263-269. (in Chinese)

定义 4 词的主题集 T. 词的主题集 T 是词 w 的主题 wt 集合,是属于词 w 的一个属性,写作 $T(w)$. 例如对于 w=足球,其主题集属性 $T(w)$=\{体育\}.

定义 5 词簇 C. 词簇 C 是 W 的一个子集. 在 C 的元素 w 含有相同主题.

定义 6 词簇的主题词 ct. $ct\in W$,对于一个词簇 C,对于所有的 $w\in C$,它们的共同主题定义为词簇的主题词.

$$\forall w_i \in C: \bigcap T(w_i) = \{ct\} \neq \emptyset, |\bigcap T(w_i)| = 1.$$

定义 7 词的关注度 wa. 词的关注度 wa 是词的一种属性,它代表了词语对一篇文章、一个句子或者一个短语的主题影响. 例如 w=“的” 的关注度为零,而 w=“原子弹”的关注度则较高.

定义 8 词簇的关注度 ca,其等同于其主题词 ct 的关注度 wa.

同义词词林[10]本身为一个类义词典,其中含有大量的分类信息. 其同一行的词语要么词义相同(有的词义十分接近),要么词义有很强的相关性. 对于同一行的词语其含有共同的主题,所以对同一行的词可以把它们聚成一个词簇. 特别的对于一些同段(包括多行)的词语,其各词仍然含有相同的主题,故可以将该段聚成一个词簇.

对所有的词簇,选取其中最有代表性的词作为词簇的主题词. 通过词林和人工判别,形成了一个基于词林的社会知识词簇集合 C_s,以下简称词簇集合.

在词簇集合 C_s 中,有些关注度较低的词簇则会被删除. 删除后剩余的词簇集合即为候选词簇集合 cC_s,而 cC_s 的主题词的集合称之为候选主题词集 cT_s.

1.2 算 法

对于短语 P,如果含有候选主题词集 cT_s 中的元素,即 $\exists w|w\in P\bigcap w\in cT_s$,则认为 w 为短语 P 所要表达的主题. 对于某些短语其含有候选主题词集的元素可能不止一个,此时认为短语包含了多个主题. 综上,对应短语 P 的主题应该是一个主题词集合.

定义 9 短语的主题词集合 Pt,对于词 w,如果 w 是短语的主题,则 $w\in Pt$.

基于上述思想,可以得到 KWPSE.

算法:KWPSE
输入:短语分词集合
输出:短语主题集合 Pt
$Pt=\emptyset$
对每个 $w\in P$ do
begin
if $w\in cT_s$ then
$Pt=Pt\cup\{w\}$
End

这种算法效率高,时间复杂度低. 但如上文所述,词语的多样化表示导致相同主题的短语表现形式不一致,P 中主题表现形式 $w\notin cT_s$. 对于此类的短语 P 利用 KWPSE 无法获取其主题信息.

举例说明如下:

给定分词后的短语 P 如下:(福建警官学院),

P=\{“福建”,“警官”,“学院”\}.

在 cT_s 中不存在这 3 个词汇. 利用 KWPSE 无法找到该短语的分类.

而事实上,学院的同义词“学校”$\in cT_s$,且“福建警官学院”的主题应为“学校”.

2 基于《知网》词相似度的短语主题抽取算法

2.1 抽取原理

KWPSE 简单地使用了 cT_s,而未考虑候选词簇集合 cC_s 所具有的类义结构含有的分类信息. 将所有的候选词簇集合 cC_s 作为训练样本集,将短语的主题抽取归约成短语主题分类. 本文利用改进的 KNN 算法进行短语主题分类.

对于某个词簇 C_i,$w(w\in C_i)$和其主题词 ct_i 存在着较大的相似度.

定义 10 P 的可能主题. 记 $w(w\in P)$对应的主题词 ct_i 为 P 的可能主题.

定义 11 备选主题词集. 短语 P 中所有可能的主题集合为短语备选主题词集,记为 $AT(P_i)=\{at_1, at_2, \cdots\}$.

本文采用对于一个短语 P,将备选词集 $AT(P)$看成是候选主题词集中距离短语主题最近的 K 个样本.

定义 12 词的相似度. $\mathrm{Sim}(w_i, w_j)$表示两个词 w_iw_j 之间的相似度.

对于一个词 w,设它所在的词簇集合为 $C(w)$,对于 $C(w)$有

$$C(w) = \{C \mid \forall C \in cC_s \bigcap w \in C\},$$

则词 w 对应的主题词集合为 $CT(w)=\{ct|\forall ct:ct$是 C 的主题词$\bigcap C\in C(w)\}$.

$\mathrm{Sim}(w, ct_i)(ct_i\in CT(w))$的值越高,则 $w\in P$ 和 $ct_i\in P$ 的关联度越高,所以我们可以把 $\mathrm{Sim}(w, ct_i)$当做 P 与 ct_i 的距离权重,即如果 $\mathrm{Sim}(w, ct_i)$越高则说明 P 与 ct_i 的距离越近.

定义 13 主题的影响度. 定义一个主题 ct 对应短

语 P 的影响度为 $I(ct,P)$.

综上所述,基于改进的 KNN 算法可以得到计算 $I(ct,P)$的方法.

对于所有的 $w_i \in P$,计算所有的 $ct_j \in CT(w_i)$与 w_i 的相似度 $\mathrm{Sim}(w_i,ct_j)$.

对于 $at_i \in AT(P)$,其影响度计算方式为

$$I(at_i,P)=\sum_{\forall j:at_i\in CT(w_j)}\mathrm{Sim}(w_j,at_i),$$

则 P 的最大影响度主题为

$$ct=ct:\max\{I(ct,P)(ct\mid ct\in AT(P))\},$$

而对于某些短语 P,其可能存在多个主题,在得到 P 的最大影响度主题 ct 后,将其他的候选主题 ct_i 与 ct 相比较,如果满足下列公式则认为 ct_i 也可能是 P 的主题.

$$\frac{I(ct,P)-I(ct_iP)}{I(ct,P)}<\alpha, \tag{1}$$

其中 α 为可接受参数,表示在允许的范围内接受 ct_i 作为 P 的主题,反应了多主题短语占所有短语的比例.本文取值为 0.03.

将符合条件的 ct_i 和 ct 合并后得到短语的主题词集合 Pt.

2.2　算法实现

基于《知网》词相似度的短语主题抽取算法(word similarity based on hownet phrase subject extraction algorithm,WSPSE)的具体实现流程如下

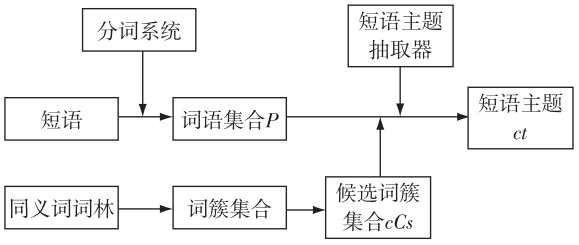

图 1　短语主题抽取算法

Fig. 1　Phrase subject extraction algorithm

本文采用中国科学院研发的 NLPIR 分词系统,该系统分词速度快,准确率高[11].

"我爱北京天安门"分词结果如例子所示,去掉分词性标注后的结果在下一行给出.

例:我/r　爱/v　北京/ns　天安门/ns./w

我爱北京天安门.

目前中文词的相似度计算有基于同义词词林和基于《知网》的词语相似度计算.考虑到《知网》的词语信息比同义词词林的含义更加完备,故在计算词的相似度时本文采用的是刘群和李素建的方法计算词的相似度[12-13].

对于两个汉语词语 W_1 和 W_2,如果 W_1 有 n 个义项(概念):$S_{11},S_{12},\cdots,S_{1n}$,$W_2$ 有 m 个义项(概念):$S_{21},S_{22},\cdots,S_{2m}$,我们规定,$W_1$ 和 W_2 的相似度为 W_1 和 W_2 对应的各个概念的相似度之最大值,也就是说:

$$\mathrm{Sim}(W_1,W_2)=\max_{i=1\ldots n,j=1\ldots m}\mathrm{Sim}(S_{1i},S_{2j}).$$

两个义原在这个层次体系中的路径距离为 d,可以得到这两个义原之间的语义距离:

$$\mathrm{Sim}(p_1,p_2)=\frac{\alpha}{d+\alpha},$$

其 p_1 和 p_2 表示两个义原,d 是 p_1、p_2 在义原层次体系结构中的路径长度,α 是可调节参数.α 的含义是当相似度为 0.5 时的词语距离值.

对于实词概念的语义表达式,将其分成 4 个部分:

第一独立义原描述式:将两个概念的这一部分的相似度记为$\mathrm{Sim}_1(S_1,S_2)$;

其他独立义原描述式:语义表达式中除第一独立义原以外的所有其他独立义原(或具体词),将两个概念的这一部分的相似度记为$\mathrm{Sim}_2(S_1,S_2)$;

关系义原描述式:语义表达式中所有的用关系义原描述式,将两个概念的这一部分的相似度记为$\mathrm{Sim}_3(S_1,S_2)$;

符号义原描述式:语义表达式中所有的用符号义原描述式,将两个概念的这一部分的相似度记为$\mathrm{Sim}_4(S_1,S_2)$.

于是,两个概念语义表达式的整体相似度记为:

$$\mathrm{Sim}(S_1,S_2)=\sum_{i=1}^{4}\beta_i\mathrm{Sim}_i(S_1,S_2),$$

其中,$\beta_i(1\leqslant i\leqslant 4)$是可调节的参数,且有:$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$.

根据文献[14],取 $\beta_1=0.5$,$\beta_2=0.2$,$\beta_3=0.17$,$\beta_4=0.13$.

计算词的相似的算法如下:

算法:词语相似度计算

输入:词 w_1 和 w_2

输出:$\mathrm{Sim}(w_1,w_2)$

从《知网》中获取 w_1 和 w_2 的概念集合$C(w_1)$和$C(w_2)$

对任意 $c_{1i}\in C(w_1)$

begin

对任意 $c_{2i}\in C(w_2)$

begin

计算 c_{1i}和c_{2i}的义原集合,并归类.

计算 $\mathrm{Sim}(c_{1i},c_{2i})=\sum_{j=1}\beta_j*\mathrm{Sim}_j(S1,S2)$

end

end

返回 $\mathrm{Max}\{\mathrm{Sim}(c_{1i},c_{2i})\mid c_{1i}\in C(w_1),c_{2i}\in C(w_2)\}$

基于上述,算法 WSPSE 如下:

算法:WSPSE 算法

输入:短语词集合 P

输出:短语主题集合 Pt

$AT(P)=\emptyset$

$Pt=\emptyset$

对每个 $w\in P$ do

begin

对 w 计算 $CT(w)$

对每个 $ct\in CT(w)$

begin

计算 $\mathrm{Sim}(w,ct)$

if $ct\in AT(P)$ then

$AT(P)[ct].\mathrm{Value}+=\mathrm{Sim}(w,ct)$

else

$AT(P)=AT(P)\bigcup\{ct\}$

$AT(P)[ct].\mathrm{Value}=\mathrm{Sim}(w,ct)$

end

end

$ct=ct:\max\{AT(P)[ct].\mathrm{Value}\ (ct\mid ct\in AT(P))\}$

对每个 $ct_i\in AT(P)$

begin

if $\dfrac{AT(P)[ct].\mathrm{Value}-AT(P)[ct_i].\mathrm{Value}}{AT(P)[ct].\mathrm{Value}}<0.03$ then

$Pt=Pt\bigcup\{ct_i\}$

end

返回 Pt

举例说明如下:

给定分词后的短语 P 如下:(福建警官学院),

$P=\{$"福建","警官","学院"$\}$.

根据词簇集合可以求出 P 的候选主题词集

$AT(P)=\{$"行政区划","军官","警察","学校"$\}$,

其中 $w_1=$"福建",$CT(w_1)=\{$"行政区划"$\}$,$\mathrm{Sim}(w_1,$"行政区别"$)=0.3$.

$w_2=$"警官",$CT(w_2)=\{$"军官","警察"$\}$,$\mathrm{Sim}(w_2,$"军官"$)=0.952$,$\mathrm{Sim}(w_2,$"警察"$)=0.933$.

$w_3=$"学院",其 $CT(w_3)=\{$"学校"$\}$,$\mathrm{Sim}(w_3,$"学校"$)=1$.

得到影响度 $I(ct,P)(ct\mid ct\in AT(P))$ 的集合为 $\{$"0.3","0.952","0.933","1"$\}$.

所以 P 的最大影响度主题 $ct=$"学校",而其他主题词与其的比值不满足公式(1),故认为其的主题为 $\{$"学校"$\}$.

3 基于加权主题词的短语主题抽取算法(WKWPSE)

3.1 算法思想

短语中还有词性信息和位置信息也是短语主题抽取所参考的信息,而 WSPSE 算法没有考虑这部分的信息.

定义 14 词的权重. 对于所有的 $w\in P$,定义 $\mathrm{Weight}(w)$ 为对应词 w 的权重.

根据不同的研究方向,对于不同的词性信息和位置信息给予不同的权重.

3.2 动名词类主题权值算法

本文研究短语主题的目的是用于搜索意图判断,故针对的短语为百度的搜索热词集合. 通过对这些短语的研究,发现这些短语大部分具有与偏正短语(如,XX 学校,XX 国家等)、动宾短语(如,学习 C 语言等)、主谓短语(如 XX 是、XX 怎么样)相同的结构.

本文抽取的短语主题主要针对动名词类主题,故认为关注度较高的词簇往往都是名词或者动词. 形成候选词簇集合 cC_s 的词簇都为名词或者动词.

基于上述考虑,名词和动词是需要重点考虑的词簇,故给予一定的权重值. 而形容词和副词则给予一个较低的权重,然后每个名词和动词的权重再加上用来修饰它的形容词或者副词的权重,这样所有的需要考虑的名词和动词都有了一定的权重.

根据语言书写习惯,较长的定语后置,而较短的定语前置. 本文研究的短语集合一般较短,我们认为名词或动词之前出现的定语都是用于修饰该动词或名词. 例如,"最美的大学"短语,"最"和"美"都是用于修饰后面的大学.

基于上述分析,本文针对动名词类的加权算法如下:

算法:WKWPSE 算法

输入:短语分局集合 P

输出:$W(w)$ 对每个 $w\in P$

对每个 $w\in P$,初始化 $W(w)=0$

对任意 $w\in P$ do

begin

通过 ICTCLAS 计算 $\mathrm{Ch}(w)$

保存 $\mathrm{Od}(w)$

end

对从大到小的 $\mathrm{Od}(w)$

begin

如果 Ch(w)是名词或动词
$W(w)=W(w)+1$
Lastvorn$=w$
如果 Ch(w)是副词或形容词
W(Lastvorn)$=W$(Lastvorn)$+0.5$
end
对任意 $w\in P$ 返回 $W(w)$

3.3　算法实现

基于上述想法,对 WSPSE 进行加权改进后形成了 WKWPSE.

算法:WKWPSE
输入:短语分局集合 P
输出:短语主题集合 Pt
$AT(P)=\varnothing$
$Pt=\varnothing$
对任意 $w\in P$ 计算 $W(w)$
对任意 $w\in P$
begin
计算 $CT(w)$
对任意 $ct\in CT(w)$
begin
计算 Sim(w,ct)
if $ct\in AT(P)$ then
$AT(P)[ct]$.Value$+=$Sim(w,ct)$\times W(w)$
else
$AT(P)=AT(P)\cup\{ct\}$
$AT(P)[ct]$.Value$=$Sim(w,ct)$\times W(w)$
end
end
$ct=ct$;max$\{AT(P)[ct]$.Value$(ct|ct\in AT(P))\}$
对任意 $ct_i\in AT(P)$
begin
if $\dfrac{AT(P)[ct].\text{Value}-AT(P)[ct_i].\text{Value}}{AT(P)[ct].\text{Value}}<0.03$ then
$Pt=Pt\cup\{ct_i\}$　　//包含 ct
end
返回 Pt

举例说明如下:

给定分词后的短语 P 如下:(厦门制服),

$P=${“厦门”,“制服”}.

根据词簇集合可以求出 P 的候选主题词集.

两个词都为名词,故其权重都为 1.而“厦门”也作为“制服”的定语,故“制服”的权重再加 0.5.由此得出两个词的权重 W(“厦门”)$=1$,W(“制服”)$=1.5$.

$AT(P)=${“城市”,“衣服”},

其中 $w_1=$“厦门”,其 $CT(w_1)=${“城市”},Sim(w_1,“城市”)$=0.57$.

$w_2=$“制服”,其 $CT(w_2)=${“衣服”},Sim(w_2,“衣服”)$=0.44$.

得到影响度 $I(ct,P)(ct|ct\in AT(P))$的集合为{“0.57”,“0.44”}.

乘以相应的权重后,P 的最大影响度主题 $ct=$“衣服”.

而其他主题词与其的比值不满足公式(1),故认为其的主题为{“衣服”}.

4　实验结果

4.1　实验数据

本次实验采用的数据是从百度搜索引擎上截取的关于“学校”、“疾病”、“衣服”、“工厂”、“商店”、“戏剧”、“乐器”、“书籍”、“婴儿”这 9 个主题的用户热门搜索短语1 198个.利用人工对这1 198个短语进行主题提取,其中学校相关 402 个,商店相关 146 个,疾病相关 134 个,衣服相关 105 个,工厂相关 69 个,戏剧相关 50 个,乐器相关 37 个,书籍相关 117 个,婴儿相关 138 个.

将文献[2]中的算法用于短语主题抽取,且将提取出的关键词在词簇中寻找主题用于表示主题,记为词频算法.

利用上述 3 种算法和词频算法分别对这1 198个短语进行主题提取.分类效果评估指标使用常用的查准率、查全率以及 $F1$ 测试值.

查准率$=$主题抽取的正确短语数/主题抽取属于该主题的短语数,

查全率$=$主题抽取的正确短语数/属于该主题的短语数,

$$F1=\frac{\text{查准率}\times\text{查全率}\times 2}{\text{查准率}+\text{查全率}}.$$

4.2　实验结果分析和比较

从结果(表 1)可以看出,词频算法直接应用于短语主题抽取,虽然其查准率较高,但查全率较低,基本与随机从短语选择主题的概率一致,故词频算法无法直接应用于短语主题抽取.而利用 WSPSE 和 WKWPSE 质量较好,且对于大多数主题的结果来说,WKWPSE 对 WSPSE 有所改进.对于乐器主题的短语,由于专有名词较多,分词词库中收录的名词并非十分全面,故其分词效果不佳,导致结果较差.而对于戏剧主题短语,其戏剧名实时更新,无法完全收录词库,不仅在分词时效果不佳,在基于同义词词林基础上形成的词簇集合也无法识别戏剧名,只能将戏剧名拆分识别

表 1 4 种算法实验结果对比表

Tab. 1 The contrast table of the result of four algorithms

算法	参数	学校	商店	疾病	衣服	工厂	戏剧	乐器	书籍	婴儿	综合
词频算法	查全率	6.47	40.30	12.38	8.70	3.42	28.00	32.43	4.27	69.57	19.28
	查准率	96.30	100.00	76.47	75.00	83.33	93.33	92.31	83.33	95.05	93.52
	*F*1	12.12	57.45	21.31	57.45	6.58	43.08	48.00	8.13	80.33	31.97
KWPSE	查全率	7.21	1.37	2.24	0.95	1.45	0.00	2.70	3.42	3.62	3.84
	查准率	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00	95.65
	*F*1	13.46	0.00	4.38	1.89	2.86	0.00	5.26	6.61	6.99	7.38
WSPSE	查全率	79.10	84.25	67.16	77.14	50.72	58.00	83.78	90.60	95.65	78.88
	查准率	99.69	99.19	100.00	98.78	100.00	100.00	93.94	99.07	97.78	99.06
	*F*1	88.21	91.11	80.36	86.63	67.31	73.42	88.57	94.64	96.70	87.83
WKWPSE	查全率	86.82	92.47	70.90	82.86	81.16	52.00	59.46	88.03	91.30	83.39
	查准率	100.00	99.26	100.00	98.86	100.00	100.00	100.00	99.04	100.00	99.70
	*F*1	92.94	95.74	82.97	90.16	89.60	68.42	74.58	93.21	95.45	90.82

导致效果较差.

4.3 多类短语实验结果和分析

从百度搜索引擎上截取部分具有多主题信息的短语,利用 WSPSE 和 WKWPSE 进行主题抽取,其实验结果如表 2.

表 2 多主题短语实验结果

Tab. 2 The result of multi-subject phrases

多主题短语	WSPSE 结果	WKWPSE 结果
经典爱情文章	爱情 文章	文章 爱情
儿童游戏	儿童 游戏	游戏 儿童
河池学院武术协会	武术 团体	武术 学校
短文两篇翻译	文章 翻译	翻译 文章
初生婴儿早期教育	婴儿 教育	教育 婴儿
安徽省立儿童医院	儿童 医院	医院 儿童
宝宝湿疹治疗	婴儿 疾病	疾病 婴儿
老年人疾病	老人 疾病	疾病 老人
教育类书籍	教育 书籍	书籍 教育
鲁迅的作品集	名人 书籍	名人 书籍

从表 2 的结果可以看出多主题短语实验结果基本符合人们主观的分主题结果,而多主题的主题抽取很大程度上依赖于社会知识词簇集合. 如果社会知识词簇集合不包含该主题信息,如上述短语中,如果“鲁迅”无法被社会知识词簇集合识别,则无法得到上述结果,只能得到“书籍”这一结果.

5 总 结

对搜索引擎的主要输入源短语进行主题提取可以更好地对用户的搜索意图进行判断,将短语抽象出主题可以将用户输入的查询表层字符信息上升到主题层面,采取不同的主题形式来形式化地表示查询背后的搜索意图,从多个角度理解查询意图,并基于查询意图提供新颖的搜索服务与搜索模式.

本文提出了对短语主题提取的算法,其中 WSPSE 和 WKWPSE 实际上是基于语义的主题提取算法. 本文实验使用的 1 198 个短语是当前热度比较高的,且具有很强的代表性,这表明本文提出的算法对短语主题提取具有积极的推进作用.

参考文献:

[1] Luhn H P. A statistical approach to mechanized encoding and searching of literary information[J]. IBM Journal of Research and Development, 1957, 1(4): 309-317.

[2] Luhn H P. The automatic creation of literature abstract [J]. IBM Journal of Research and Development, 1958, 2 (2): 159-165.

[3] Edmundson H P, Oswald V A, Wyllys R E. Automatic indexing and abstracting of contents of documents[R]. Los Angeles: Planning Research Corp, 1959.

[4] Stevens M E. Automatic indexing: a state-of-the-art report[EB/OL]. [2014-10-29]. http://digital.library.unt.

edu/ark:/67531/metadc171070/.

[5] 马颖华,王永成,苏贵洋,等.一种基于字同现频率的汉语文本主题抽取方法[J].计算机研究与发展,2003,6:874-878.

[6] 杨洁,季铎,蔡东风,等.基于联合权重的多文档关键词抽取技术[J].中文信息学报,2008,22(6):75-79.

[7] 李素建,王厚峰,俞士汶,等.关键词自动标引的最大熵模型应用研究[J].计算机学报,2004,27(9):1192-1197.

[8] 李鹏,王斌,石志伟,等.Tag-TextRank:一种基于 Tag 的网页关键词抽取方法[J].计算机研究与发展,2012,11:2344-2351.

[9] 宋巍.基于主题的查询意图识别研究[D].哈尔滨:哈尔滨工业大学,2013.

[10] Che W X,Li Z H,Liu T. LTP:a Chinese language technology platform[C]// Proceedings of the Coling 2010: Demonstrations. Beijing,China:[s. n.],2010:13-16.

[11] 中国科学院. ICTCLAS 汉语分词系统[EB/OL].[2010-12-21]. http://www.ictclas.org.

[12] 董振东,董强.知网(HowNet)[EB/OL].[1999-06-01]. http://www.keenage.com.

[13] 刘群,李素建.基于《知网》的词汇语义相似度计算[J].中文计算语言学,2002,7(2):59-76.

Phrase Subject Extraction Based on Synonyms and HowNet

ZENG Cong,ZHANG Dong-zhan*

(School of Information Science and Engineering,Xiamen University,Xiamen 361005,China)

Abstract: Key word phrase subject extraction algorithm (KWPSE), which is based on the judgment whether phrases include the topic words is constructed. On the basis of KWPSE,by using a WordsSet of social knowledge as classified information,the word similarity as distance weight,and the improved KNN method the word similarity based on HowNet phrase subject extraction algorithm (WSPSE) is presented. Finally,on this basis of WSPSE and with the addition of the weight to the words' position that is based on Chinese custom,the WKWPSE algorithm is proposed. The average recall rates reach 78.88% and 83.39%,and average precision rates increase to 99.06% and 99.70%.

Key words: phrase subject;subject extraction;synonyms;HowNet