MB-SinglePass:基于组合相似度的微博话题检测

周 刚[1,2] 邹鸿程[2] 熊小兵[2] 黄永忠[2]

(软件开发环境国家重点实验室 北京 100191)[1] (信息工程大学信息工程学院 郑州 450002)[2]

摘 要 话题检测技术在传统媒体的研究中取得了较好的效果。探讨了针对微博类的新型媒体短文本对象话题检测技术的优化及性能评价。基于微博中联系人存在的关注和粉丝等结构化信息、帖子之间转发评论等内在关联关系,提出了针对微博的MB-SinglePass话题检测算法。该算法除了考虑微博上述特点之外,还针对短文本特征稀疏的问题,利用同义词典,引入了微博特征扩展技术,丰富了特征信息。同时,针对单一使用余弦相似度、雅各比相似度和语义相似度的不足,采用了组合相似度策略。相较传统算法,MB-SinglePass算法在新浪微博实测数据集上取得了更好的性能。另外,针对相似度策略的对照实验说明采用组合相似度的效果优于单一相似度。

关键词 微博,SinglePass,话题检测,文本相似度,同义词扩展

中图法分类号 TP391 **文献标识码** A

MB-SinglePass:Microblog Topic Detection Based on Combined Similarity

ZHOU Gang[1,2] ZOU Hong-cheng[2] XIONG Xiao-bing[2] HUANG Yong-zhong[2]

(State Key Laboratory of Software Development Environment, Beijing 100191, China)[1]

(Information Engineering Institute, Information Engineering University, Zhengzhou 450002, China)[2]

Abstract Topic detection achieves quite good result in the traditional media research. This paper discussed the refiness and performance evaluation of the topic detection technique in the new kind of medias such as microblog, proposed the MB-SinglePass topic detection algorithm on the basis of the structured information such as the relationships of attentions and fans between contacts, the inner connection relationships such as forwarding and comment between posts. Beside considering the above microblog characteristics, MB-SinglePass introduces the characteristics extension technique in order to enrich characteristics information. At the same time, the paper used the combined similarity aiming at the shortage of singly utilizing the Jaccard similarity coefficient, cosine based similarity and semantic similarity. Compared with the traditional algorithms, MB-SinglePass shows better performance on the actual dataset of sina microblog. Additionally, experiment according to the similarity strategy reveals better result by using combined similarity than singular similariy.

Keywords Microblog, SinglePass, Topic detection, Text similarity, Synonyms extension

1 引言

自微博问世以来,其因实时移动、简短快捷和互动方便等特点迅速风靡全球。例如,国外著名的微博网站twitter用户突破1.7亿,每天锐推数量超过1.3亿条。国内首个微博网站新浪微博也以极快的速度在发展。微博上通过相互关注、转发评论等行为形成了一个庞大的社交网络。微博文本限制在140个字以内,一般即时发布,因此微博语言呈现出个性化、符号化、口语化、非规范化的特点,也产生了很多新鲜词汇,比如"神马"、"给力"等,从某种程度上引领了一种社会文化。

根据TDT(话题检测与追踪,Topic Detection and Tracking)组织的任务划分,话题检测技术属于TDT的一项子任务[1]。话题检测又称主题检测,其目的是对一系列连续的文本流自动检测出不同的话题。它本质是一种无指导的聚类。向量空间模型(VSM)是文本形式化的主要模型之一[1]。通过将文本映射到多维空间,使得对文本相似比较计算有了可行性。话题检测算法常用的方法有agglomerative算法[2]、增量聚类算法[3]、增量K-means算法[4]、基于主题模型算法[5]等。相似度计算的方法常采用余弦夹角、雅各比公式、OKAPI公式、Clarity、Tanimoto、Hellinger公式等[6,7]。随着自然语言处理技术的发展,基于语义相似度的相似度模型也开始流行[8]。这些相似度计算方法各有利弊。

话题检测与追踪技术在传统研究对象如新闻、论坛和博

到稿日期:2011-12-07 返修日期:2012-03-06 本文受软件开发环境国家重点实验室开放课题(SKLSDE-2011KF-06),国家高技术研究发展(863)计划(2009AA043303)资助。

周 刚(1974—),男,博士,副教授,CCF会员,主要研究方向为海量信息处理、云计算、社会网络分析,E-mail:gzhougzhou@gmail.com;邹鸿程(1985—),男,硕士生,主要研究方向为网络数据挖掘;熊小兵(1985—),男,博士生,主要研究方向为网络数据挖掘;黄永忠(1968—),男,博士,教授,博士生导师,主要研究方向为分布式系统、云计算、海量信息处理。

客等文本较长的媒体中取得了较好的效果。在微博平台上，由于帖子文本内容较短，使得产生特征稀疏问题，给话题检测与技术追踪技术的应用带来了较大的挑战。但是，微博平台自身具有的特点为话题检测与追踪技术研究带来了新的思路，这也是本文研究的出发点。首先，随着自然语言处理相关技术的发展，同义词扩展技术得到发展[9]，并且取得了较好的效果。本文利用该技术对微博文本进行特征扩展，尝试解决特征稀疏的问题。其次，微博平台具备一些传统媒体不具有的特性，如用户之间关注行为、用户相互转发评论行为等。这些结构化特征为研究话题检测与追踪技术带来了新的机遇。在微博主题检测中可以充分利用这些结构化信息辅助判断，以提高主题检测的性能。再有，如前所述，相似度策略各有优缺点，于是，本文提出了组合相似度的方法。它融合了余弦相似度、雅各比相似度和语义相似度 3 种策略。这是基于以下 3 点考虑：第一，余弦相似度采用的是对特征进行 TF-IDF 数量化后再进行计算的策略；而雅各比相似度仅仅利用特征数量统计结果进行相似度计算；语义相似度则是基于知识库的相似度计算方法。这 3 种方法角度各异，有着本质的区别。第二，余弦相似度和雅各比相似度在计算特征相互不重合的两个文本时，会直接得出其相似度为 0，而实际在微博上这两条帖子可能讨论一个相关的事件，从而造成误判。如果引入语义相似度策略，则可以减少其中的误差。第三，由于知网等知识库收录的登录词汇有限，语义相似度存在相似性判断不够精确的问题。因此，融合 3 种方法可以从多个角度综合地刻画文本的相似特征，避免单一方法的不足，从而提高文本相似度计算的准确性。

本文的主要工作如下：

1. 提出了组合相似度策略，充分利用不同相似度策略的优点，同时规避其缺点。

2. 针对微博特点，提出了 MB-SinglePass 算法，结合帖子转发回复的信息，对帖子的类别进行判断。同时，针对微博特征稀疏问题，利用《同义词林》对关键词进行特征扩展，丰富了微博文本的特征。

3. 通过在新浪微博实际数据集上测试，MB-SinglePass 算法显示出较好的话题检测性能。

2 MB-SinglePass 话题检测算法

2.1 基于同义词扩展的帖子特征模型

帖子形式化的一般步骤是对帖子进行分词，去除停用词，统计并计算词的 TF-IDF 权重。经过形式化，每个帖子与一个多维向量相对应，见式(1)：

$$pos_i \sim (V_{i1}, V_{i2}, \cdots, V_{ij}, V_{ik}) \tag{1}$$

式中，k 表示词空间大小，即向量空间的维度；V_{ij} 表示帖子 i 中词 j 的 TF-IDF 权重，根据式(2)计算得出[10]：

$$\begin{cases} V_{ij} = tf_{ij} * idf_j \\ tf_{ij} = \dfrac{c_{ij}}{M_i}, idf = \lg\left(\dfrac{N}{n} + 0.01\right) \end{cases} \tag{2}$$

式中，tf_{ij} 表示词 j 在帖子 i 中的词频，c_{ij} 表示词 j 在帖子 i 出现的次数，M_i 表示帖子 i 的总词数，N 表示总的帖子数，n 表示出现词 i 的帖子数。0.01 是为了避免出现 0 值而设定的一个常量。

由于微博帖子较为简短，特征本身就比较稀疏，为了尽可能地利用特征信息，本文借助《同义词林》对主题的关键词向量进行特征扩展，使关键词特征数量扩大 2 倍以上。本文设定各扩展特征的权重与其来源特征的权重相同。

2.2 主题模型

在基于向量空间模型的主题模型中，主题是由一簇帖子组成的。所以，主题向量的物理意义可以理解为主题里各帖子向量的叠加向量的平均，如图 1 所示。具体的主题向量生成过程是：

1. 对各个主题簇内的帖子进行分词，去除停用词，去重并计算 TF-IDF 权重形成初始主题向量。

2. 对经过步骤 1 之后形成的关键词进行特征扩展，形成更丰富的特征向量。

3. 对主题内扩展后的文本向量进行叠加平均，得到的向量即为主题向量。

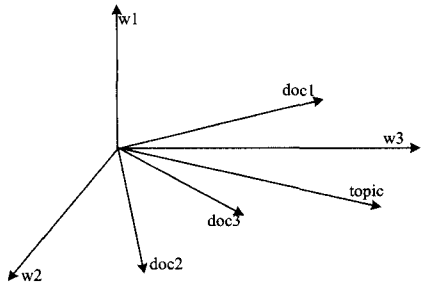

图 1 主题向量示意图

2.3 组合相似度策略

2.3.1 知网语义相似度

《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[11]。

《知网》中有两个主要的概念："概念"与"义原"。"概念"是对词汇语义的一种描述。每一个词可以用一个或多个概念来描述。"概念"在知网里是用"知识表示语言"来描述的。"义原"是描述"概念"的最小意义单位，也是"知识表示语言"所用的"词汇"。与其他语义词典不同，《知网》并不是简单地将所有的"概念"归结到一个树状的概念层次体系中，而是试图用一系列的"义原"来对每一个"概念"进行描述。

本文使用的语义相似度方法参考了刘群、李素建的论文——基于《知网》的词汇语义相似度计算[11]，其中，针对知网收录词，语义相似度计算包括以下几种情况：

1. 实词与虚词相似度为 0。

2. 虚词概念只有{句法义原}或者{关系义原}两种描述方式，因此，两个虚词概念之间的相似度等于对应句法义原或关系义原之间的相似度，见式(3)：

$$\mathrm{sim}(p_1, p_2) = \frac{\alpha}{\alpha + d} \tag{3}$$

式中，d 表示义原 p_1，p_2 在知网义原层次体系中的路径长度，α 是一个可调节参数。

3. 实词概念可以用 4 种不同义项类型来描述，第一种是第一基本义原描述式，是对实词最重要的一种描述；第二种是除第一基本义原之外的其他基本义原描述式，是一个义原集合；第三种是关系义原描述；第四种符号义原描述。一般地，两个实词概念之间的相似度可由式(4)来计算：

$$\mathrm{sim}(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} \mathrm{sim}_j(S_1, S_2) \tag{4}$$

式中，$sim_j(S_1,S_2)(j=1,2,3,4)$表示 S_1 和 S_2 两个概念对应上述 4 种不同义项的两两相似度。β_i 是可调节的参数，且满足：$\beta_1+\beta_2+\beta_3+\beta_4=1$；$\beta_1>\beta_2>\beta_3>\beta_4$。这种参数设置方法的意义是使主要义项的相似度对次要义项的相似度进行制约，维持主要义项相似度对整体相似度的主要贡献地位。例如，如果次要义项相似度高，主要义项相似度低，那么整体相似度也不会受到次要义项的太大影响。

针对知网未收录词，本文采用以下方法进行处理：

1. 对于人名

如果两个人名姓氏部分相同，则认为两者具有高相似度，如潘石屹、老潘和潘总；

如果两个人名除姓氏部分相同，则认为其具有较高相似度，如吴菊萍和菊萍；

对于外籍人名，如果存在部分相同匹配，则认为其具有较高相似度，这主要考虑了外来词音译带来的误差，如本拉登、拉丹和本拉丹。

2. 对于地名

针对知网只收录市一级及以上行政区域地名的限制，手工编制各地区简称及下属县、区地名表。对于市级以下行政级别之间的相似度，通过分别将其提升到对应的知网收录的市级以上行政区域进行相似度计算。如郑州金水区与北京朝阳区相似度，可以用知网收录词郑州与北京的相似度来近似。同时，规定属于同一市的各级行政地名之间相似度为 1。如金水区与郑州市。

3. 对于机构组织

对于机构组织存在简称的情况，如果两个机构名称通过了部分相同匹配，则不加区别，即相似度为 1。如“住建部”与“住房与城乡建设部”相似度为 1。

4. 其他情况

除以上情况之外，如果两个知网未收录词相同，则相似度为 1，否则为 0。

2.3.2 组合相似度

常用的相似度计算方法有余弦相似度、雅各比相似度[12]、语义相似度[8]等。具体地：

余弦相似度的定义为：

$$sim_{cos}(pos_i,pos_j)=\frac{\sum_{n=1}^{k}V_{in}\times V_{jn}}{(\sum_{n=1}^{k}V_{in}^2)(\sum_{n=1}^{k}V_{jn}^2)} \tag{5}$$

雅各比相似度的定义为：

$$sim_{jac}(pos_i,pos_j)=\frac{|pos_i\cap pos_j|}{|pos_i\cup pos_j|} \tag{6}$$

式中，$|pos_i\cap pos_j|$表示帖子 i 和帖子 j 的公共特征数量，$|pos_i\cup pos_j|$表示帖子 i 和帖子 j 的不同特征数量。

借鉴信息融合中的顺序加权思想，语义相似度的定义如下：

设帖子 i 含有 m 个关键词，对应的权重为 $v_{i1},v_{i2},\cdots,v_{im}$，见式(7)：

$$pos_i\sim[word_{i1}/v_{i1},word_{i2}/v_{i2},\cdots,word_{im}/v_{im}] \tag{7}$$

帖子 j 含有 n 个关键词，对应的权重为 $v_{j1},v_{j2},\cdots,v_{jn}$，见式(8)：

$$pos_j\sim[word_{j1}/v_{j1},word_{j2}/v_{j2},\cdots,word_{jn}/v_{jn}] \tag{8}$$

则帖子 i 与帖子 j 的语义相似度归一化表示见式(9)(由于对称性，不妨假设 $m\geqslant n$，反之亦然)：

$$\begin{cases}sim_{yuyi}(pos_i,pos_j)=\\ \dfrac{v_{i1}\,simsen1(word_{i1},pos_j)+\cdots+v_{im}\,simsen1(word_{im},pos_j)}{\sum_{r=1}^{m}v_{ir}}\\ simsen1(word_{ir},pos_j)=\max\{simsen2(word_{ir},word_{jl})\end{cases} \tag{9}$$

式中，$l=1,2,\cdots,n$；$r=1,2,\cdots,m$；$simsen2(word_i,word_j)$表示词 i 和词 j 的相似度；$simsen1(word_i,pos_j)$表示词 i 和帖子 j 的相似度。

如引言所述，这 3 种相似度的思想和角度不同，而且仅单一使用一种相似度都存在一些不足。本文基于以上认识，提出了组合相似度策略，见式(10)：

$$sim_{comb}(pos_i,pos_j)=sim_{cos}(pos_i,pos_j)\times\alpha+sim_{jac}(pos_i,pos_j)\times\beta+sim_{yuyi}(pos_i,pos_j)\times\gamma \tag{10}$$

式中，α,β,γ 表示加权系数，反映了 3 种不同相似度对总体相似度的贡献大小。

2.4 主题模型的更新策略

随着帖子依次流入，帖子数量发生变化的对应主题的向量也相应地得到更新。但是，新帖子的加入可能是正确的一种划分，即相关帖子；也可能是不正确的划分，即伪相关帖子。如果仅仅使用新帖子集合的叠加向量作为来更新主题向量，会存在较大的误差。因此，本文在算法过程中，保留各主题的历史向量，并将历史向量加入主题更新模型。各历史向量对最新主题模型的贡献随着时间的推移和更新次数的增加逐渐衰减。基于以上思想，给出主题更新模型如下：不妨假定当前是第 n 次更新主题模型，当有新帖子划分到话题时，该主题模型按式(11)更新；若没有新帖子加入话题，主题向量与前一次的主题向量相同。

$$T_n=\frac{\sum_{i=1}^{n-1}\alpha_i T_i+\alpha_n T_{ntemp}}{\sum_{i=1}^{n}\alpha_i} \tag{11}$$

式中，α_i 表示主题向量的加权系数，它是一个与更新次数有关的参数，定义为式(12)：

$$\alpha_i=\frac{1}{n-i+1} \tag{12}$$

T_i 表示第 i 次更新的主题向量，T_{ntemp} 表示第 n 次主题临时向量，它是由新帖向量与 T_{n-1} 叠加而成。即

$$T_{ntemp}=V+T_{n-1} \tag{13}$$

式中，V 是第 n 次新来的帖子向量。

2.5 MB-SinglePass 算法流程

基于向量空间模型的话题检测常采用的方法是双阈值算法，又称 SinglePass 算法[1]，其主要思想是通过设立聚类阈值 v_c 和创新阈值 v_n 两个门限来判别将帖子归入某个类别或创建一个新类。由于微博的特殊性，本文将微博帖子的转发评论关系、联系人的相互关注关系等辅助信息加入算法辅助考虑，提出了 MB-SinglePass 算法。MB-SinglePass 算法过程如图 2 所示，具体描述如下：

Step1 初始化处理。主要包括对帖子进行分词、去除停用词、去重、计算 TF-IDF 权重等处理。

Step2 特征扩展。采用《同义词林》词典对帖子特征进行扩展。扩展特征的权重与原特征的权重相同。

Step3 依次读取到来的帖子向量，若是第一条帖子，则创建新话题。如果不是，转 Step4。

Step4　判断新帖子与当前已判断的帖子是否存在转发评论关系，如果有，则直接归入相应话题，并更新保存话题向量；如果没有，则计算新帖子与各话题向量的最大相似度。如果最大相似度大于聚类阈值 v_c，则归入相应话题，并更新保存话题向量，转 Step6；否则，转入 Step5。

Step5　判断最大相似度与创新阈值 v_n 之间的关系，如果其值小于创新阈值 v_n，则创建新话题，转 Step6；否则，判断新贴子与对应最大相似度话题内帖子的发贴人是否存在关注和好友关系，如果存在，则归入相应话题类，并更新保存话题向量，转 Step6；如果不存在，帖子归入相应话题类，但不更新话题向量，以上一次话题向量当作本次话题向量，同样保存当前的主题向量记录，以便用于下一次更新话题向量。转 Step6。

Step6　判断是否为最后一条帖子，若不是，转入 Step3；若是，结束算法。

输入：按发帖时间顺序依次读入帖子 $pos_1, pos_2, \cdots, pos_n$

输出：各个帖子话题簇 $C_1, C_2, \cdots$

1. {预处理：分词、去停用词、去重、赋权、同义词扩展}
2. count=1;
3. $C_1 \leftarrow pos_1$;
4. save $T[1]_1$;
5. repeat
6. count++;
7. if(pos_{count}与已判断帖子 t 存在转发评论关系)then
8. pos_{count}归入帖子 t 所在话题簇 a;
9. update and save $T[a]_{count}$;
10. update and save $T[other]_{count}$;
11. end
12. else
13. [maxsim,ClusterNo]=CalMaxSim(pos_{count},T[]);//计算最大相似度 maxsim 及对应簇号 ClusterNo
14. if(masimx>v_c)then
15. pos_{count}归入 ClusterNo 簇
16. update and save $T[ClusterNo]_{count}$;
17. update and save $T[other]_{count}$;
18. end
19. else if(maxsim<v_n)then
20. create cluster new and save T[new];
21. update and save $T[other]_{count}$;
22. end
23. else if(pos_{count}与 ClusterNo 簇内已判断帖子发贴人存在关注或粉丝关系)then
24. pos_{count}归入 ClusterNo 簇;
25. update and save T[ClusterNo];
26. update and save $T[other]_{count}$;
27. end
28. else
29. pos_{count}归入 ClusterNo 簇;
30. $T[ClusterNo]_{count}=T[ClusterNo]_{count-1}$;
31. save $T[ClusterNo]_{count}$;
32. update and save $T[other]_{count}$;
33. end
34. end
35. end
36. end
37. until(count>n)

图 2　MB-SinglePass 算法

为了更加清晰直观地表示各步骤之间的关系，图 3 给出了 MB-SinglePass 算法流程图。

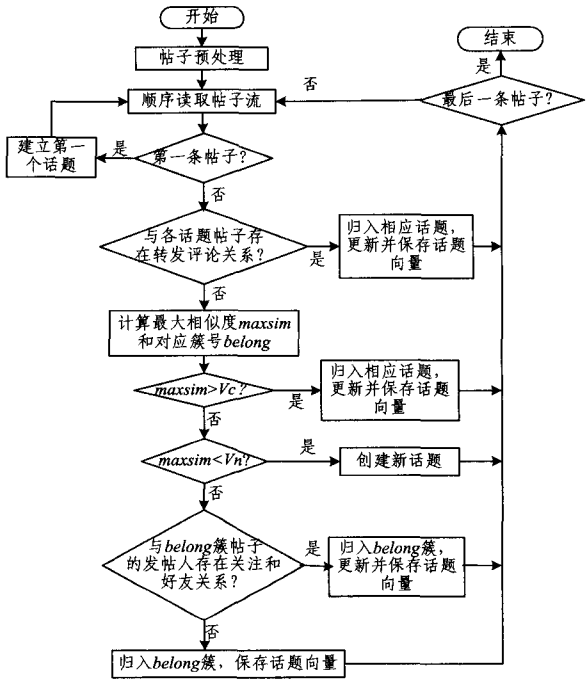

图 3　MB-SinglePass 算法流程图

2.6　评价方法

本文评价方法采用 NIST 建立的话题检测与追踪的评价方法，主要的性能指标包括准确率、召回率、漏检率、误检率、准确率和召回率的综合指标 *F1-measure* 以及模型指标 $Norm(C_{Det})$。各指标计算公式如下[1]：

$$\text{准确率 } P=\frac{a}{a+b} \tag{14}$$

$$\text{召回率 } R=\frac{a}{a+c} \tag{15}$$

$$\text{漏检率 } P_{miss}=\frac{c}{a+c} \tag{16}$$

$$\text{误检率 } P_{FA}=\frac{b}{b+d} \tag{17}$$

$$F1\text{-}measure=\frac{2PR}{P+R} \tag{18}$$

$$Norm(C_{Det})=\frac{(C_{miss}\cdot P_{miss}\cdot P_{target}+C_{FA}\cdot P_{FA}\cdot P_{non-target})}{\min(C_{miss}\cdot P_{target},C_{FA}\cdot P_{non\text{-}target})} \tag{19}$$

式中，a 为检测到的帖子数，b 为检测到的不相关帖子数，c 为检测到的相关帖子，d 为检测到的不相关帖子。一般地，C_{miss} 取为 1，C_{FA} 取 0.1，P_{target} 取 0.02，$P_{non\text{-}target}$ 取 0.98。

3　实验及结果分析

3.1　数据采集及预处理

本文以新浪微博为实验平台，采集了清华百年校庆、房地产调控、个税起征点上调、盈江地震、药家鑫事件、杭州最美妈妈、北京地铁事故、本拉登遇袭身亡、日本核辐射、威廉王子大婚等 10 个热点话题共 108981 条帖子数据，以及发贴时间、相应的发贴人及其帖子之间是否存在转发评论等信息。为了更好地测试算法，从每个话题中人工地挑选了 100 条帖子质量好的数据共计 1000 条帖子进行测试。转发评论关系主要根据新浪微博里的转发标志“//@”来进行识别，即标志前的帖

子是对标志后帖子的转发评论。为了获得发贴人之间的关注和粉丝关系，选定实验帖子后，通过正则表达式析取出发贴人ID号，进一步获得并建立所有发贴人的关注和粉丝列表，以便在算法过程中匹配发贴人关系。

针对待测试的帖子，采用分词工具对其进行分词、对分词结果去除停用词、去重，并计算每个词的 TF-IDF 权重，最后采用同义词典对帖子特征进行扩展，并赋予权重。中文分词工具采用的是中科院 ICTCLAS2011 提供的 JNI 接口，同义词典采用的是哈工大信息检索研究室的同义词词林扩展版。去除停用词是在构建一个停用词典的基础上进行的，该词典还包括微博常用表情符号等，共有 487 个词。基于此词典，匹配并过滤帖子中存在的停用词。

为了确定组合相似度的加权系数，分别对 $\alpha+\beta+\gamma=1$ 且 $\alpha,\beta,\gamma>0$ 的 36 种取值可能情况（精度为 0.1）测试 MB-SinglePass 算法的 *F1-measure* 值。实验表明，当 $(\alpha,\beta,\gamma)=(0.3,0.3,0.4)$ 时，*F1-measure* 取得最大值 0.8432。因此，本文组合相似度的参数 α,β,γ 取值分别取为 0.3、0.3、0.4。在实际应用中，根据实际问题的不同，最佳的加权系数可能是不同的。

3.2 实验 1

为了测试 MB-SinglePass 算法的性能，本文设计了一组对比试验。对照算法也是话题检测常用的增量聚类和增量 K-means 算法。增量聚类算法只采用单一阈值策略，算法实际类似于二值分类问题；增量 K-means 算法按帖子数量设置窗口长度，值设为 50，共两个窗口。实验采用 10 个话题的平均性能作为相应算法的性能指标。实验结果如图 4 所示。

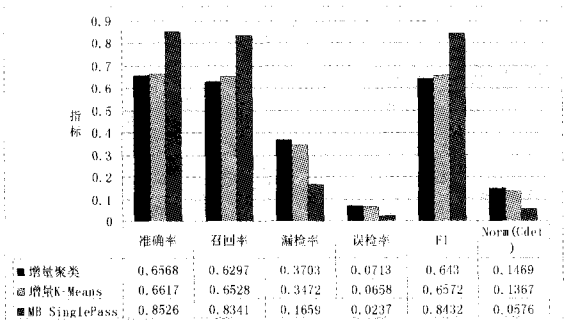

图 4 算法性能比较结果

从以上实验数据可以得到以下结论：首先，相较其他两个算法，MB-SinglePass 算法在各项指标上都有较大的提高，性能提高将近 30%。这主要是由于 MB-SinglePass 算法考虑了微博的结构化信息，而且采用了特征扩展技术，丰富了特征信息，同时组合相似度策略充分吸取了不同相似度策略的优点也是一个重要原因。其次，引入微博的相关信息，可以较大地提高检测算法的性能。在增量聚类和增量 K-means 算法中，由于没有引入微博的各种结构化信息，加上微博的特征稀疏性，使得算法无法得到满意的性能。因此，在微博话题检测算法中引入结构化信息虽然增加了复杂度，但却是必要的。

3.3 实验 2

为了测试采用组合相似度策略的效果。基于同一实验数据，本文采用 MB-SinglePass 算法，对余弦相似度、语义相似度和组合相似度 3 种不同相似度策略进行了对照比较。单一使用雅各比相似度效果较为一般，本文未进行探讨。实验的数据及参数与实验 1 相同。实验结果如图 5 所示。

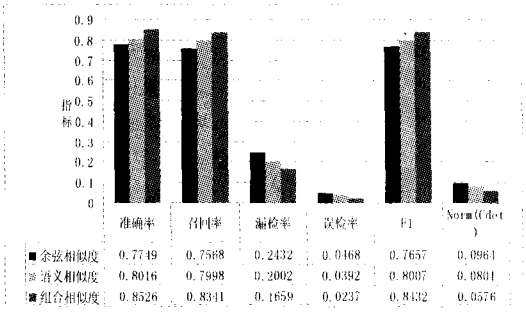

图 5 不同相似度策略下 MB-SinglePass 算法的性能比较

在同一实验参数和检测算法情况下，通过测试比较，可以看到使用组合相似度策略结果性能指标明显优于余弦相似度和语义相似度，最多性能可以提高 10%。实验充分说明了组合相似度可以更好地刻画相似度特征。此外，单一使用语义相似度比单一使用余弦相似度性能更好，性能提高 5%以上。这主要是因为余弦相似度没有考虑到语义信息，而语义信息在短文本相似度比较中是很重要的因素。

结束语 本文针对微博短文本提出了 MB-SinglePass 算法。与基于传统媒体对象的话题检测算法相比，MB-SinglePass 算法考虑了微博平台本身的特点，即微博结构化信息，包括发贴人之间存在的关注关系、帖子的转发评论关系等。同时 MB-SinglePass 算法采用了组合相似度和同义词扩展技术，更准确地刻画了相似性和丰富了帖子的特征信息。通过与传统话题检测算法增量 K-means、增量聚类的比较显示，MB-SinglePass 算法在性能上有较大的提高。下一步的研究可以在其他微博平台上进一步测试，同时对其他算法进行测试，甚至可以在话题模型上进一步探讨，以使微博的话题检测性能得到进一步的提高。

参 考 文 献

[1] 洪宇，张宇，等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，21(6)

[2] 闪四清，陈茵，程雁. 数据挖掘—概念、模型、方法和算法[M]. 北京：清华大学出版社，2003：114-116

[3] Allan J, Carbonell J, Doddington G. Topic Detection and Tracking Pilot Study: Final Report[C]// Proceeding of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco, 1998: 194-218

[4] Yamron J P, Knecht S, van Mulbregt P. Dragon's Tracking and Detection Systems for the TDT2000 Evaluation[C]// Proceeding of Topic Detection and Tracking Workshop. Washington, USA, 2000: 75-80

[5] 贾自艳，何清，等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展，2004，41(7)：1273-1280

[6] Brants T, Chen F, Farahat A. A system for new event detection [C]// Proceedings of the 26th SIGIR Conference on Research and Development in Information Retrieval . 2003

[7] 张晓艳，王挺. 话题检测与追踪技术研究[J]. 计算机研究与探索，2009，3(4)：347-357

[8] 席耀一，林琛，李弼程，等. 基于语义相似度的论坛话题追踪方法[J]. 计算机应用，2011，31(1)：93-96

[9] 徐建民，唐万生. 基于查询术语同义词的扩展信念网络检索模型[J]. 计算机工程，2007，33(10)：28-30

[10] 孙胜平. 中文微博客热点话题检测与追踪技术研究[D]. 北京：北京交通大学，2011

[11] 刘群，李素建. 基于《知网》的词汇语义相似度计算[C]// 第三届汉语词汇语义学研讨会. 台北，2002：59-76

[12] Rangrej A, et al. Comparative Study of Clustering Techniques for Short Text Documents[C]// WWW2011. Hyderabad, India, 2011