第35卷 第20期 计 算 机 工 程 2009年10月
Vol.35 No.20 Computer Engineering October 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)20—0184—03 文献标识码：A 中图分类号：TP311.52

基于 WFC 和 MI 的主题句提取方法

薛扣英[1]，原 盛[1]，张心严[2]
(1. 西安交通大学电子与信息工程学院，西安 710049；2. 西安交通大学软件学院，西安 710049)

摘 要：提出一种基于加权模糊聚类(WFC)和互信息(MI)的主题句提取方法，使主题句尽可能全面覆盖全文主题的同时，缩减自身的冗余，以提高摘要效率，采用加权模糊聚类的方法对文本句子进行分类，对在同一类中的句子使用比较互信息的方法进行排名处理，从而获得高质量的摘要。实验结果表明，与传统聚类方法比较，该方法的正确率提高约15%，可以达到约70%的精确度，并在阅读摘要时能够基本正确地获取文本信息。

关键词：主题句；加权模糊聚类；互信息

Topic Sentence Extraction Method Based on Weight Fuzzy Clustering and Mutual Information

XUE Kou-ying[1], YUAN Sheng[1], ZHANG Xin-yan[2]
(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049;
2. School of Software, Xi'an Jiaotong University, Xi'an 710049)

【Abstract】 A topic sentence extraction method based on Weight Fuzzy Clustering(WFC) and Mutual Information(MI) is proposed, which is to cover more topics and lower the redundant information of the text. The abstract efficiency is promoted. Using WFC method, the sentences is classified. Sentences in each cluster is ranked by MI values. High qualified abstact is obtained. Experimental results show that, compared with former clustering method, this method can improve the precision by nearly 15%, and has about 70% accuracy. It can get text information correctly.

【Key words】 topic sentence; Weight Fuzzy Clustering(WFC); Mutual Information(MI)

1 概述

自动摘要分为2类：自动抽取和自动提取。自动抽取着眼于内容完全由从输入文档中拷贝的内容组成。而自动提取的内容则包含了一些输入文档中没有表达的东西。从描述和计算的角度来看，文本理解在很大程度上仍然是一个没有解决的问题，因此，不需要理解文本的抽取方法对自动文本摘要的生成非常很重要。

常用的抽取方法有2种：有指导的方法和无指导的方法。在无指导的摘要中，聚类方法被广泛采用，比较目前较流行的使用在文摘上的2种聚类方法：K-means聚类[1]和FCM(模糊聚类法)[2]，发现这2种算法对敏感点处理依赖于初始的聚类中心并且产生空簇现象。例如：假定要聚类的向量具有10维属性，其中只有4维属性对聚类最相关，这4维属性具有相似值的向量在10维属性空间中却有可能距离最远，此时由10维属性等同作用的相似度量则引起了误导。克服此问题的有效方法就是为每一个属性加特征权参数，让不同的属性在聚类中起不同的作用[3]。本文在模糊聚类算法的基础上给每一维属性加权值，形成加权模糊聚类法(Weight Fuzzy Clustering Method, WFCM)，并对空簇现象进行处理。

聚类结果的每一类中包含了相似度较大的句子，可以认为这些句子的内容较为接近，它们叙述的是文章主旨的同一侧面。因此，需要对类内的句子进行比较排名，选择最能代表本簇的句子集合。对于句子之间的关系比较，很多文献使用的是句子之间相似度计算，也有使用pageRank思想对句子进行排名，然而这些算法都只考虑句子之间的相似性，忽略了句子之间的相异性，产生大量冗余，而对一些主题的忽略，使得摘要不具有完整性。考虑到同一个簇内，挑选最大限度能代表该簇信息的句子，达到冗余度最小，也要覆盖率尽可能的高的思想。本文把互信息(Mutual Information, MI)的思想引入到句子排名中，考虑同一个簇内句子之间的相异性，尽可能覆盖多个主题。

2 基于向量空间模型的句子与文档表示方法

2.1 基于句子的向量空间模型

人们使用名词主要是使用其理性义，发挥其指称功能。所以，在选择关键词的时候更多的考虑名词，其中包括词库内有的名词以及未登陆划分出来的名词信息。每个文档 D 的句子特征表示为向量：$\boldsymbol{S}=(f_{i1},f_{i2},\cdots)$，对句子向量进行特征提取得到向量空间模型 $\boldsymbol{VSM}_i=(f_{i1},f_{i2},\cdots)$，包括如下信息：

(1)句子长度信息；
(2)句子与关键字列表的相似性信息；
(3)句子与标题相似性信息；
(4)句子位置特征信息；
(5)句子平均tf*idf信息；
(6)句子平均名词信息。

基金项目：中科院国际合作伙伴计划基金资助项目(2F05N01)
作者简介：薛扣英(1984－)，女，硕士研究生，主研方向：自然语言处理，数据挖掘；原 盛，讲师、博士研究生；张心严，硕士研究生
收稿日期：2009-05-16 **E-mail：**xkouying@gmail.com

2.2 基于词形和词序相似度计算的句子与标题相似度计算

标题包括非常重要的关键词信息，所以，需要考虑句子与标题之间的相似度才能够更好的对句子进行聚类。词形相似度表示 2 个句子包含的相同词语数目占所有不同词语的总数目的比例，把分词得到的所有词语和符号作为特征项。$Same(A,B)$ 表示句子 A，B 中相同特征项的个数，当一个特征项在 A，B 中出现次数不同时，以出现次数少的计数，$len(A)$，$len(B)$ 分别表示句子 A，B 包含的特征项的个数，则句子 A，B 的词形相似度为

$$WordSim(A,B)=2\times\frac{Same(A,B)}{len(A)+len(B)},\ 0\le WordSim(A,B)\le 1$$

词序相似度表示逆序数占在 2 个句子中都出现但只出现一次的词语的数目的比例，$OnceWS(A,B)$ 表示句子 A，B 都出现且只出现一次的特征项的集合，$\boldsymbol{Pfirst}(A,B)$ 表示 $OnceWS(A,B)$ 的特征项在 A 中的位置序号构成的向量，$\boldsymbol{Psecond}(A,B)$ 表示 $\boldsymbol{Pfirst}(A,B)$ 中的分量按对应特征项在 B 中的次序排序生成的向量，$RevOrd(A,B)$ 表示 $Psecond(A,B)$ 中各相邻分量的逆序数，则句子 A，B 的词序相似度 $OrdSim(A,B)$ 为

$$OrdSim(A,B)=\begin{cases}1-RevOrd(A,B)/\\(|OnceWS(A,B)|-1) & |OnceWS(A,B)|>1\\1 & |OnceWS(A,B)|=1\\0 & |OnceWS(A,B)|=0\end{cases}$$

句子 A，B 的相似度可表示为

$$Sim(A,B)=\lambda_1\times WordSim(A,B)+\lambda_2\times OrdSim(A,B)$$

其中，λ_1，λ_2 为常数，且 $\lambda_1+\lambda_2=1$，对于 λ_1，λ_2 的选择，本系统根据其在句子相似度中的不同作用，设定 $\lambda_1=0.8$，$\lambda_2=0.2$。

2.3 基于经验特征的位置加权表示

位置特征信息加权表示，句子的位置在句子分类中占有重要的地位，加权规则：

$$location_i=(f_{i1},f_{i2},f_{i3},f_{i4},f_{i5})$$

因为主题句都是陈述句，所以若句子是陈述句，则 $f_{i1}=1$，若是第 1 段和最后一段的第 1 句或是最后一句，则 $f_{i2}=1$，若是第 1 段和最后一段的其他句子，则 $f_{i3}=1$，若是其他段的第 1 句或者是最后一句，则 $f_{i4}=1$，若是其他段的其他句子，则 $f_{i5}=1$。

根据经验，加权公式如下：

$$W_i=(3f_{i1}+3f_{i2}+2f_{i3}+2f_{i4}+f_{i5})/5$$

2.4 聚类个数的确定

聚类算法一个重要的方面是确定聚类个数，本文通过用户指定摘要比例进行聚类个数确定，如果用户不指定，默认的摘要比例为 20%，在后续有实验数据进行经验认证。聚类个数计算方法如下：

$$k=\lfloor(Text\_length\times percent)/average\_sen\_length\rfloor \quad (1)$$

根据经验可得很短或很长的句子都不太可能出现在摘要中，而句子长度适中的句子比较有可能出现在摘要中，所以，通过式(1)得出初始的聚类个数是可行的。

3 基于加权的模糊聚类算法

聚类算法分很多类，包括层次聚类、平面聚类等。而在自动文摘中常用 K-means 算法和 FCM(模糊 c 均值算法)。

聚类的目标函数为

$$J(U,V)=\sum_{i=1}^{n}\sum_{k=1}^{c}(\mu_{ik})^m(d_{ik})^2 \quad (2)$$

其中，μ_{ik} 是第 k 个元素属于第 X_i 个簇的隶属度；$J(U,V)$ 表示各类中样本到聚类中心的加权距离平方和，$m\in(0,\infty)$，m 为加权值；$V=[v_k\mid v_k$表示第k个聚类中心，$k\in(1,\infty)]$；d_{ik} 一般为欧拉距离 $d_{ik}=\sqrt{|x_i-v_k|^2}$。

3.1 K-means 聚类方法

当 μ_{ik} 取如下值时，为硬聚类。

$$\mu_{ik}=\mu_{X_i}(x_k)=\begin{cases}1 & x_k\in X_i\\0 & x_k\notin X_i\end{cases}$$

最优化目标函数式(2)即为 K-means 聚类方法，通过迭代式(2)，引导寻求最优解。

3.2 模糊 c 均值

模糊聚类考虑隶属度 μ 在 [0,1] 之间：

$$\begin{cases}\mu_{ik}=\mu_{X_i}(x_k)=\{\mu_{ik}\mid\mu_{ik}\in[0,1]\}\\\sum\limits_{0\le i\le mk}\mu_{ik}=1\end{cases}$$

根据聚类的标准，现要求最小化目标函数式(1)，即 min($J(U,V)$)。

根据拉格朗日乘数法求解可得迭代式如下：

$$\mu_{ik}^{(b)}=\left\{\sum_{j=1}^{c}\left[\left(\frac{d_{ik}^{(b)}}{d_{jk}^{(b)}}\right)^{\frac{2}{m-1}}\right]\right\}^{-1} \quad (3)$$

$$V_i^{(b+1)}=\frac{\sum\limits_{k=1}^{n}\left(\mu_{ik}^{(b)}\right)^m\cdot x_k}{\sum\limits_{k=1}^{n}\left(\mu_{ik}^{(b)}\right)^m},\quad i=1,2,\cdots,c \quad (4)$$

在 K-means 算法上提出了改进，在每次计算中给出一个模糊值，这样一个向量属于某个聚类就有一定的概率程度，选取隶属度最高的为最后聚类。

3.3 加权模糊聚类方法

在 FCM 算法的基础上加在每一维空间的权重，在式(2)的基础上改进得到如下 WFCM 的目标函数：

$$\min J_w(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\sum_{j=1}^{s}w_j\cdot d(x_{kj},v_{ij}) \quad (5)$$

其中，w 的更新函数如下[4]：

$$diff\_hit=\sum_{j=1}^{R}\frac{|x_i-h_j|}{\max(X)-\min(X)}$$

$$diff\_miss=\sum_{l\ne class(x_i)}\frac{P(l)}{1-P(class(x_i))}\sum_{j=1}^{R}\frac{|x_i-m_{lj}|}{\max(X)-\min(X)}$$

$$w=w-\frac{diff\_hit}{R}+\frac{diff\_miss}{R} \quad (6)$$

其算法步骤如下：

步骤 1 随机从数据集中选取 K 个点作为初始聚类中心，初始化 ***weight*** 向量。

步骤 2 计算各个样本到聚类中心的距离，把样本归到离它最近的那个聚类中心所在的类。

步骤 3 计算新形成聚类的数据对象的平均值得到新的聚类中心，如果相邻两次的聚类没有任何变化，说明样本调整结束，跳至步骤 5，否则跳至步骤 4。

步骤 4 按照式(3)更新隶属度向量，按照式(4)更新聚类中心，按照式(6)更新权重，跳至步骤 2。

步骤 5 聚类结束，比较每个句子的隶属度，选择最大的得到聚类结果。

4 基于互信息的簇内句子排名

信息论中互信息的定义如下[3]：

$$I(X,Y)=\sum_{0\le i\le count} p(x_i,y)\mathrm{lb}\left(\frac{p(x_i,y)}{p(x_i)p(y)}\right) \tag{7}$$

如果 X 和 Y 是相互独立的，则 $I(X,Y)=0$，如果依赖性越高，则 $I(X,Y)$ 值越大。

对于同一聚类内句子的排名，考虑 MI，即考虑句子之间的相异性，可以区分句子之间对于关键字的包含程度来决定该句子的排名。对句子中的关键字进行 MI 计算，根据 MI 的值进行排名，该值的物理含义是主题与该句子的相异程度，完全相异为 0，分值越高，表示离主题越近，可以覆盖的主题越多。聚类考虑的是相似性，而 MI 考虑相异性，两者结合得到比较好的效果。

当式(7)中 X 表示关键词，Y 表示句子时，可简化为

$$I(X,Y)=\sum_{0\le i\le count} p(x_i,y)\mathrm{lb}\left(\frac{p(x_i,y)}{p(x_i)p(y)}\right) \tag{8}$$

其中，$p(x_i,y)$ 可以简化为第 i 个关键词在 y 句子中出现的频率；$p(x_i)$ 为第 i 个关键词在文本中出现的频率；$p(y)$ 为 y 句子的长度占整个文章长度的比例。

在同样的句子长度下，如果 MI 值大的话，表明句子覆盖的主题更多，句子长度长，关键词多，不一定 MI 的值越大，这符合正常的假设。

常用的句子排名方式是对句子 ***VSM*** 向量进行加权平均，根据经验，参考其他文献加权值向量为 ***weight*** =(0.2,0.5,0.4, 0.5,0.3,0.2)，每一项分别表示句子长度、句子关键字比例、句子与标题的相似度、句子位置信息、句子平均 tf*idf、句子名词信息。

5 测试与分析

自动摘要系统的测试目前没有统一的评价标准，常用的评价方法是把摘要系统生成的摘要与相同文章的人工专家摘要进行对比。首先定义量化的评价标准：一篇人工专家摘要所含的全部句子总数记为 Mf ；系统生成的摘要中包含的全部句子总数记为 Sf 。其中，系统生成的摘要中所包含的与人工专家摘要相一致的句子总数记为 Cf 。以下 3 个量化指标就可以作为对两种类型摘要对比测试结果的量化评价标准：

(1)召回率(Recall)

$$R=\frac{Cf}{Mf}$$

(2)精度(Precision)

$$P=\frac{Cf}{Sf}$$

(3)F 指数(F-measure)

$$F=2\times\frac{R\times P}{R+P}$$

本文从新浪网上随机下载了不同类型的文章 100 篇，其中包括经济、体育、娱乐以及目前关注的地震、奥运会等，文章类别具有普遍性。实验中首先按照不同类别，不同摘要比例测试数据，表 1 给出使用本文 WFCM 和 MI 方法下的摘要结果，可以明显看出摘要比例越高，摘要效果越好，可是如果摘要比例太高，当文章篇幅较多时，摘要过长而不满足用户需求，所以，本文最后采用提取 20%的摘要比例，这个比例在测试中能够得到较好的结果，而且也是在日常经验中被认可的比较好的摘要比例，表 2 给出在 20%的摘要比例下不同摘要方法的结果。

表 1　不同摘要比例不同文章类型的 WFCM 方法的摘要结果

文章类型	文章数目	摘要比例/(%)	基于互信息排名比较的方法/(%)			基于位置权重排名比较的方法/(%)		
			P	*R*	*F*	*P*	*R*	*F*
政治类	30	10	68.08	60.29	63.95	51.09	52.01	51.55
		20	73.21	68.09	70.56	54.07	53.04	53.55
		30	74.28	71.90	73.07	54.05	55.08	54.56
经济类	30	10	66.26	60.87	63.45	49.80	51.21	50.50
		20	70.58	66.28	68.36	51.02	53.07	52.02
		30	74.28	70.28	72.22	55.01	56.63	55.81
生活类	20	10	69.50	69.58	69.54	50.07	54.06	51.99
		20	72.28	73.89	73.08	51.04	56.01	53.41
		30	74.28	75.58	74.92	57.30	58.72	58.00
体育类	10	10	66.28	67.29	66.78	48.09	51.07	49.54
		20	72.28	71.15	71.71	47.09	49.08	48.06
		30	74.28	73.15	73.71	50.02	52.07	51.02
评论类	10	10	31.58	32.85	32.20	28.05	21.08	24.07
		20	35.58	29.29	32.13	27.54	25.05	26.24
		30	37.26	33.15	35.09	28.09	27.08	27.58

表 2　摘要比例为 20%的各种方法摘要结果

使用方法	基于互信息排名比较的方法/(%)			基于位置权重排名比较的方法/(%)		
	P	*R*	*F*	*P*	*R*	*F*
K-means	60.38	62.65	61.49	41.73	42.94	42.33
Fuzzy-means	67.35	70.00	68.65	45.01	46.47	45.73
weightF-means	73.77	70.16	71.87	52.93	54.15	53.50

从表 1、表 2 可以看出，WFCM 和 MI 组合的方法与其他传统的聚类方法的比较提高了近 15%的正确率，而 70%左右的精确度在阅读摘要时能够基本覆盖文本内容而且用户能够比较清楚地了解文本信息。同时，该方法比传统的聚类方法以及传统的句子加权方法都有比较好的性能，而把两者结合起来的新方法具有更好的性能。但本文方法对评论性的文章效果不是很好，原因在于评论性的文章结构性比较散，主题信息不能很容易分类，导致各种基于聚类的方法效果都不是很理想。

6 结束语

本文提出的把加权模糊聚类方法和互信息结合起来的方法达到良好的性能，但是新方法需要大量的经验统计信息进行修正，如何通过大量文本训练出各个参数的值是将要重点研究的内容。

参考文献

[1] 江开忠，李子成，顾君忠. 自动文本摘要方法[J]. 计算机工程, 2008, 34(1): 221-223.

[2] Banerjee A. A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation[J]. The Journal of Machine Learning Research, 2007, 8: 1919-1986.

[3] 李双虎，王铁洪. K-means 聚类分析算法中一个新的确定聚类个数有效性的指标[J]. 河北省科学院学报, 2003, 20(4): 200-202.

[4] 李　洁，高新波，焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报, 2006, 34(1): 89-92.

编辑　陈　文