基于关联词对动态抽取的报道关系检测技术研究

赵 华[1] 邓 攀[2] 张建伟[2]
(山东科技大学信息科学与工程学院 青岛 266510)[1]
(北京航空航天大学软件开发环境国家重点实验室 北京 100191)[2]

摘 要 报道关系检测是判断随机选取的两个新闻报道是否讨论同一话题的技术。提出了一种基于关联词对动态抽取的报道关系检测方法。关联词对是指在同一篇报道中出现的满足一定关系约束的两个单词,而关系约束是指一组特征的集合。该方法认为两篇报道中出现的相同的关联词对越多,两篇报道的相似度越大。实验证明基于关联词对动态抽取的报道关系检测方法取得了非常好的效果,从而证实了所提方法的有效性。同时,实验还表明,关系约束对该方法的成功实施起着非常重要的作用。

关键词 话题检测与跟踪,报道关系检测,关联词对,关系约束
中图法分类号 TP391 **文献标识码** A

Story Link Detection Research Based on the Dynamic Extraction of Correlative Word

ZHAO Hua[1] DENG Pan[2] ZHANG Jian-wei[2]
(College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266510, China)[1]
(National Lab of Software Development Environment, Beijing University of Aeronautics, Beijing 100191, China)[2]

Abstract Story link detection is a task of topic detection and tracking, which is to detect whether two stories are "linked", or discuss the same event. This paper proposed a story link detection method based on the dynamic extraction of correlative word. Correlative word is defined to be a pair of words that satisfy certain Relation Restriction. In this paper, relation restriction refers to a set of features. This paper explored three features, which include the initial letter, location and window. The more two stories had the same correlative words, the larger similarity they had. Experimental results showed that the story link detection method based on the dynamic extraction of correlative word performs very well, which testifies the great capabilities of this method. We also find that relation restriction is critical to the performance of this method.

Keywords Topic detection and tracking, Story link detection, Correlative word, Relation restriction

报道关系检测是话题检测与跟踪(Topic Detection and Tracking, TDT)研究的一个重要的核心技术,它是判断随机选取的两个新闻报道是否讨论同一个话题的技术。

本文认为,为了得到高性能的报道关系检测系统,必须充分挖掘报道内容中所包含的信息。受词汇共现的启发,提出了基于关联词对动态抽取的报道关系检测方法。关联词对是指在同一篇报道中出现的满足一定关系约束的两个单词,而关系约束是指一组特征的集合,当然也可以是空集。动态抽取是指关联词对是按照一定的关系约束从报道中动态地抽取出来,是动态变化的。该方法认为,如果同时满足某特定的关系约束的几个单词经常共同出现,则它们在一定程度上表达了该报道的语义信息。例如,在一篇新闻报道中,如果单词"Janpan","earthquake"成为关联词对次数比较多,则可以推断这应该是一篇是关于日本地震的新闻报道。文中通过分析指出,传统的词汇共现模型不适用于面向新闻话题的报道关系检测。

从上述的描述中可以看出,关系约束对于本文方法的成功实施至关重要,这是因为关系约束定义得越好,从报道中抽取出来的关联词对就越能体现报道的语义信息。本文在构建关系约束时选用了3个特征,分别为首字母特征、位置特征以及距离特征。实验结果证明,基于关联词对动态抽取的英语报道关系检测系统取得了理想的效果,提高了报道关系检测系统的性能。

1 相关工作介绍

目前,最成功的报道关系检测系统的基本做法是:首先将报道表示成向量空间模型,然后使用Cosine函数计算两个向量之间的相似度;最后通过阈值和相似度之间的比较做出报道是否相关的判断:如果相似度大于阈值,那么报道相关;否则报道不相关。

到稿日期:2009-07-09 返修日期:2009-09-23 本文受国家自然科学基金(60773034),山东科技发展计划项目(2008GG30001024),山东科技大学"春蕾计划"项目资助。

赵 华(1980—),女,博士,讲师,主要研究方向为话题检测与跟踪等,E-mail:doctorhuazhao@yahoo.com.cn;邓 攀(1983—),女,博士生,主要研究方向为语义网络等;张建伟(1978—),男,博士生,主要研究方向为语义网络等。

上述的基本模型中主要解决两个方面的问题：报道的表示模型以及报道之间的相似度计算。UMass 验证了多种相似度计算方法在报道关系检测系统中的性能[1]，包括余弦函数、加权和、语言模型以及交叉熵，并得出结论：Cosine 函数在报道关系检测中的性能最好。受多分类器融合思想[2]的启发，Francine Chen 等人将多种相似度计算函数结合起来用于计算报道之间的相似度，并取得了很好的效果[3]。另外，为了充分利用报道的内容信息，Ying-Ju Chen 等人将多种自然语言处理技术以及信息检索技术应用于单语及多语的报道关系检测系统中，例如报道扩展、话题切分等[4]。文献[5]则在报道关系检测研究中探索使用了动态停用词技术，取得了理想的效果。部分研究者对报道关系检测与其他 TDT 任务的之间的关系进行了研究，如与报道切分之间的关系[6]、与新事件检测之间的关系[7,8]。

词汇共现模型是基于统计方法的自然语言处理研究领域的重要模型之一。它是建立在这样一个基本假设的基础之上：如果在大规模语料中，两个词经常共同出现在文档的同一窗口单元，则认为这两个词在意义上是相互关联的。根据词汇共现模型，若某几个单词经常在同一窗口单元中共同出现，则它们在一定程度上表达了该文本的语义信息，这是因为作者通常倾向于通过在不同的句子中重复这些词来强调文本主题。

本文认为传统的词汇共现模型在报道关系检测中不适用，因为：传统的词汇共现是指在大规模语料中，两个词经常共同出现在文档的同一窗口单元。然而报道关系检测是面向新闻领域中的新闻报道的，且话题是不停地动态演化的，且通常由几个关键词汇相互区分，如人名、地名、组织机构名等。所以本文认为，基于新闻话题的这些动态特征使得基于大规模语料事先统计出来的相对静态的词汇共现不适用于面向新闻话题的报道关系检测。基于此，本文提出了动态抽取关联词对的方法。

2 报道关系检测基本模型

本文中用于报道关系检测的基本模型的框架如下：

(1) 首先对报道进行预处理操作，包括去停用词和词形还原；

(2) 将报道表示成空间向量模型，其中特征项是报道中出现的不同词，特征项的权值则是词在报道中的词频；

(3) 使用余弦函数计算报道向量之间的相似度。假设 $w_{11}, w_{12}, \cdots, w_{1n}$ 和 $w_{21}, w_{22}, \cdots, w_{2n}$ 分别为特征项 $\delta_1, \delta_2, \cdots, \delta_n$ 在报道 $S1$ 和报道 $S2$ 中的权值，则 $S1$ 和 $S2$ 的基于余弦函数的相似度 $\mathrm{Cos}(S1,S2)$ 计算公式如下：

$$\mathrm{Cos}(S1,S2)=\sum_{k=1}^{n} w_{s1k}\times w_{s2k}/\sqrt{\sum_{k=1}^{n} w_{s1k}^2\times\sum_{k=1}^{N} w_{s2k}^2} \quad (1)$$

(4) 通过阈值和相似度之间的比较做出报道是否相关的判断：如果相似度大于阈值，那么报道相关；否则报道不相关。

3 基于关联词对动态抽取的报道关系检测算法

3.1 基本概念

定义 1(关联词对) 本文中关联词对是指出现在同一篇报道中的满足一定关系约束的两个单词。为了便于描述以及方便自然语言信息处理的具体应用，可以将一个关联词对形式化为一个三元组$\langle W1, W2, RR\rangle$，其中 $W1$ 与 $W2$ 表示出现在报道中的词，而 RR 表示词对满足的关系约束(Relation Restriction，RR)。当处于某特定的关系约束下时，也可将关联词对简化为一个二元组$\langle W1, W2\rangle$。

定义 2(关系约束) 关系约束是一组特征的集合，可以形式化为 $\emptyset$ 或者$\{f1, f2, \cdots\}$，其中 $f1, f2$ 表示特征。

当 $RR=\emptyset$ 时，报道中任何一个词对都将形成一个关联词对，这样将会得到许多没有意义的、冗余的关联词对。另一方面，从相关文献中可知，新闻报道往往只是依据几个关键词或短语来区分新闻事件。为了能抽取出更加有意义的关联词对，本文在构建关系约束时中采用了以下几个特征：

(1)首字母大写特征：在英语报道中，表示名称(人名、地名及组织机构名)的单词的首字母通常都大写，而名称是区分相似话题的关键信息[9]，所以将首字母大写特征作为抽取有效关联词对的依据之一，中用 C(Capital)表示该特征。

(2)位置特征：新闻报道具有独特的头重脚轻的倒金字塔式结构，一般把最重要的事实写在报道的开头部分，所以本文认为出现在报道较前部分的单词具有较高的重要性。基于此，本文采用位置特征作为抽取有效关联词对的依据之一，本文使用 L(Location)表示该特征。

(3)距离特征：距离特征(窗口特征)是词汇共现模型中的一个重要的特征。所以我们有理由相信距离特征也是构建关系约束的一个有用的特征。本文将两个词之间的距离定义为出现在这两个词中间的词的个数。例如在句子"It reduces air pollution"中，air 和 pollution 之间的距离为 0，而 reduces 和 pollution 之间的距离为 1。本文中使用 D(Distance)表示该特征。

上述特征可以两两组合起来形成组合特征，总的来说，本文验证了以下特征的性能：

$RR=\{C=1\}$：报道中任何两个首字母大写的词组成一个关联词对。

$RR=\{D\}$：报道中位于一定距离内的任何两个词组成一个关联词对。本文中 $D=\alpha$ 表示单词之间的距离小于等于 α，其中 $0\leqslant\alpha\leqslant\omega-2$，$\omega$ 是报道的长度。

$RR=\{L\}$：位于报道的前 L 部分的单词两两组成一个关联词对，其中 $0<L\leqslant 1$。

$RR=\{C,D\}$：位于一定距离范围内的首字母大写的两个词组成一个关联词对。

$RR=\{C,L\}$：位于报道的前 L 部分的任何两个首字母大写的词组成一个关联词对。

$RR=\{D,L\}$：位于报道的前 L 部分且在一定距离范围内的任何两个词组成一个关联词对。

$RR=\{C,D,L\}$：位于报道的前 L 部分且在一定距离范围内的任何两个首字母大写的词组成一个关联词对。

定义 3(动态抽取) 本文中动态抽取是指关联词对每次都是从待比较的两篇报道中动态抽取的，即关联词对是动态变化的。这一点与词汇共现是不同的，词汇共现是从大规模语料中事先统计出来的，是相对静止的。

为了使得抽取出的关联词对更加有意义，本文在抽取关联词对之前首先对报道进行预处理操作，包括去停用词和词形还原，在词形还原过程中保留了首字母大写信息。

为了更加清楚地介绍本文中的基本概念，给出了如下的

例子：对于经过预处理后的句子："World Bank commit support Palestinian Authority face economic challenge"，在各种不同的关系约束下从该句中所抽取的关联词对如表 1 所列。值得说明的是，为了节省空间，在表 1 中将关系约束统一写在第一列，而将关联词对简化为〈W1，W2〉。

表 1　动态抽取关联词对举例

RR	关联词对
{C=1}	〈World，Bank〉；〈World，Palestinian〉；〈World，Authority〉；〈Bank，Palestinian〉；〈Bank，Authority〉；〈Palestinian，Authority〉
{D=0}	〈World，Bank〉；〈Bank，commit〉；〈commit，support〉；〈support，Palestinian〉；〈Palestinian，Authority〉；〈Authority，face〉；〈face，economic〉；〈economic，challenge〉
{L=1/3}	〈World，Bank〉；〈World，commit〉；〈Bank，commit〉
{C=1，D=0}	〈World，Bank〉；〈Palestinian，Authority〉
{C=1，L=1/3}	〈World，Bank〉
{D=0，L=1/3}	〈World，Bank〉；〈Bank，commit〉
{C=1，D=0，L=1/3}	〈World，Bank〉

此处的例子只是为了解释基本概念，在使用过程中，关联词对是从整篇新闻报道中按照一定的关系约束进行动态抽取的。另外，从表 1 中可以看出，关系约束对于方法的成功实施至关重要，这是因为关系约束定义得越好，从报道中抽取出来的关联词对就越能体现报道的语义信息。从上述的例子中也可以看出，本文提出的关联词对在建立过程中不需要背景知识，简单灵活。同时，通过第 2 节的分析可知：基于新闻话题的动态特征，使得基于大规模语料事先统计出来的相对静态的词汇共现不适用于面向新闻话题的报道关系检测。

3.2　基于关联词对的相似度计算方法

本文将动态抽取的关联词对应用于报道之间的相似度计算中。假设 $CW(S1)$ 和 $CW(S2)$ 分别表示报道 $S1$ 和报道 $S2$ 中的在某个特定的关系约束下动态抽取的关联词对的个数，而 $SameCW(S1,S2)$ 表示 $S1$ 和 $S2$ 所共有的关联词对的个数，那么基于关联词对的报道之间的相似度 $CW\mathrm{Sim}(S1,S2)$ 如下所示：

$$CW\mathrm{Sim}(S1,S2)=\frac{2\times SameCW(S1,S2)}{CW(S1)+CW(S2)} \quad (2)$$

从式(2)可以看出，两篇报道中相同的关联词对的个数越多，报道之间相似度也越大。

3.3　基于关联词对的报道关系检测算法

假设要参与检测的两个报道为 $S1$ 和 $S2$，那么基于关联词对动态抽取的报道关系检测算法的过程如下所示：

(1) 对报道 $S1$ 和报道 $S2$ 分别进行预处理，并将它们都表示成向量空间模型；

(2) 统计分别出现在报道 $S1$ 和报道 $S2$ 中的关联词对，并统计它们之间相同的关联词对；

(3) 分别使用式(1)和式(2)计算报道 $S1$ 和报道 $S2$ 的基于余弦函数的相似度以及基于关联词对的相似度；

(4) 使用式(3)计算报道 $S1$ 和报道 $S2$ 的最终相似度 $Final\mathrm{Sim}(S1,S2)$：

$$Final\mathrm{Sim}(S1,S2)=\mathrm{Cos}(S1,S2)+CW\mathrm{Sim}(S1,S2) \quad (3)$$

(5) 将最终相似度和预设的阈值进行比较，作出 $S1$ 和 $S2$ 是否相关的判定：如果最终相似度大于阈值，那么 $S1$ 和 $S2$ 相关，否则 $S1$ 和 $S2$ 不相关。

4　实验与结果分析

4.1　评价标准及语料

依据 TDT 评测标准，本文采用归一化检测开销 $(C_{Lnk})_{Norm}$ 来评测报道关系检测系统的性能，公式如下：

$$(C_{Lnk})_{Norm}=\frac{C_{Miss}\cdot P_{Miss}\cdot P_{target}+C_{FA}\cdot P_{FA}\cdot P_{\rightarrow target}}{\min(C_{Miss}\cdot P_{target},C_{FA}\cdot P_{\rightarrow target})} \quad (4)$$

式中，P_{Miss} 为系统的漏报率；P_{FA} 为系统的误报率；P_{target} 为在信息流中看到一个新话题的概率；$P_{\rightarrow target}$ 为在信息流中看到一个老话题的概率，$P_{\rightarrow target}=1-P_{target}$；$C_{Miss}$ 为漏报一个新话题的代价；C_{FA} 为误报一次的代价。

$(C_{Lnk})_{Norm}$ 越小表明系统的性能越好，理想情况下，$(C_{Lnk})_{Norm}=0$。依据 TDT 评测标准，本文实验中，C_{Miss}，C_{FA} 及 P_{target} 的值分别为 1.0，0.1，0.02。

我们利用已有的 TDT Pilot Corpus 手工建立了用于英语报道关系检测系统的评测语料。在建立过程中，使用话题 1 至话题 5 来建立训练语料，而话题 6 至话题 8 用来建立测试语料。语料的建立方法为：与同一个话题相关(标记为 YES)的两两报道组成一个相关报道对；分别与不同的话题相关(标记为 YES)的两两报道组成一个不相关报道对。

综上所述，本文用于英语报道关系检测的训练语料与评测语料的相关统计如表 2 所列。

表 2　英语报道关系检测语料

	训练语料	测试语料
相关的报道对数	1545	2644
不相关的报道对数	4126	2816
总计	5671	5460

4.2　实验设置与结果分析

4.2.1　报道关系检测基本模型性能验证

我们首先测试了报道关系检测基本模型的性能。通过在训练语料上的测试可知，报道关系检测 baseline 系统在相似度阈值等于 0.22 时取得最好的结果，其评测结果为：漏报率 0.1172，误报率 0.0046，归一化检测开销 0.1397(取得此结果的系统中报道的表示模型中不对词性加以区分)。在相似度阈值等于 0.22 的设定下，对报道关系检测 baseline 系统在评测语料上的性能进行了评测，其评测结果为：漏报率 0.0783，误报率 0.0071，归一化检测开销 0.1131。

4.2.2　基于关联词对动态抽取的报道关系检测系统性能验证

为了验证本文方法的有效性，在不同的关系约束的限定下验证了基于关联词对动态抽取的报道关系检测系统的性能，结果如表 3 所列，其中关系约束中各参数值都是通过训练语料训练得到的。

表 3　基于关联词对动态抽取的报道关系检测系统实验结果

RR	漏报率	误报率	归一化检测开销	实验名称
{C}	0.0386	0.0089	0.0821	RRisC
{D=0}	0.0609	0.0078	0.0992	RRisD
{L=1/9}	0.0753	0.0071	0.1101	RRisL
{C，D=20}	0.0257	0.0103	0.0762	RRisCD
{C，L=1/3}	0.0480	0.0082	0.0881	RRisCL
{D=0，L=1/4}	0.0643	0.0075	0.1008	RRisDL
{C，D=10，L=1/3}	0.0556	0.0075	0.0921	RRisCDL

为了更好地比较各个特征的性能，给出了表 4 和表 5。从上述 3 个表中的实验结果可以得出以下结论：

(1) 基于关联词对动态抽取的报道关系检测系统的性能在各个不同的关系约束限制下都比基本模型系统有了明显的改善，这就说明本文提出的基于关联词对动态抽取的报道关

(下转第 270 页)

[C]// Proc. of the Congress on Evolutionary Computation. Seoul, Korea, 2001: 101-106

[5] Natsuki H. Particle swarm optimization with Gaussian mutation [C]// Proc. of the Congress on Evolutionary Computation. Indianapolis, Indiana, 2003: 72-79

[6] Shi Y, Eberhart R. A modified particle swarm optimizer[C]// IEEE World Congress on Computational Intelligence. Piscata-way: IEEE Press, 1998: 69-73

[7] Zhang L P, Yu H J, Hu S X. A new approach to improve particle swarm optimization[C]// Lecture Notes in Computer Science. Chicago: Springer Verlag, 2003: 134-139

[8] Chen G M, Huang X B, Jia J Y, et al. Natural exponential inertia weight strategy in particle swarm optimization[C]// Proc. of 6th Congress on Intelligent Control and Automation. Dalian: IEEE Press, 2006: 3672-3675

[9] 何庆元,韩传久. 带有扰动项的改进粒子群算法[J]. 计算机工程与应用,2007,43(7):84-86

[10] 黄辉先,陈资滨. 一种改进的粒子群优化算法[J]. 系统仿真学报,2007,19(21):4922-4925

[11] 任子晖,王坚. 一种动态改变惯性权重的自适应粒子群算法[J]. 计算机科学,2009,36(2):227-229

[12] 倪庆剑,邢汉承,张志政,等. 动态概率粒子群优化模型及实验分析[J]. 计算机科学,2009,36(2):222-226

[13] 郝柏林. 从抛物线谈起——混沌动力学引论[M]. 上海:上海科技教育出版社,1993

[14] 舒斯特 H. 混沌学引论[M]. 成都:四川教育出版社,1994

[15] 黄贤英,张丽芳. 基于粒子群优化的模糊聚类算法[J]. 重庆工学院学报:自然科学版,2008,22(11):120-123

(上接第 239 页)

系检测方法非常成功。

表 4 D 取不同值时 RRisD 与 RRisCD 的性能

Experiment Name / Value of D	RRisD			RRisCD		
	漏报率	误报率	归一化检测开销	漏报率	误报率	归一化检测开销
0	0.0609	0.0078	0.0992	0.0250	0.0295	0.1694
1	0.0681	0.0071	0.1029	0.0321	0.0217	0.1383
2	0.0715	0.0067	0.1045	0.0329	0.0167	0.1147
3	0.0745	0.0067	0.1076	0.0352	0.0135	0.1013
4	0.0749	0.0071	0.1097	0.0352	0.0131	0.0996
5	0.0741	0.0071	0.1089	0.0340	0.0121	0.0932
6	0.0749	0.0071	0.1097	0.0337	0.0114	0.0893
7	0.0745	0.0071	0.1093	0.0325	0.0110	0.0865
8	0.0730	0.0071	0.1078	0.0329	0.0110	0.0868
9	0.0734	0.0078	0.1117	0.0318	0.0110	0.0857
10	0.0738	0.0075	0.1103	0.0321	0.0107	0.0843

表 5 L 取不同值时 RRisL 与 RRisCL 的性能

Experiment Name / Value of L	RRisL			RRisCL		
	漏报率	误报率	归一化检测开销	漏报率	误报率	归一化检测开销
1/4	0.0696	0.0085	0.1114	0.0688	0.0050	0.0932
1/5	0.0715	0.0078	0.1098	0.0707	0.0050	0.0951
1/6	0.0734	0.0075	0.1099	0.0734	0.0046	0.096
1/7	0.0734	0.0078	0.1117	0.0775	0.0050	0.1019
1/8	0.0745	0.0075	0.111	0.0802	0.0046	0.1028
1/9	0.0753	0.0071	0.1101	0.0817	0.0046	0.1043
1/10	0.0760	0.0071	0.1108	0.0836	0.0046	0.1062

(2) 从实验 RRisC,RRisD 和 RRisL 的结果可以看出,本文使用的几个特征都取得了很好的效果。从 RRisD 与 RRisCD,以及 RRisL 与 RRisCL 的结果对比中可以发现,首字母大写特征非常成功,这是因为在英语报道中,表示名称(人名、地名及组织机构名)的单词的首字母通常都大写,而名称是区分相似话题的关键信息。

(3) 从结果可以看到,与首字母大写特征以及距离特征相比,位置特征的性能稍微弱些,我们认为这是由于话题的动态演化特性引起的[10]。

结束语 本文认为为了得到高性能的报道关系检测系统,必须充分挖掘报道内容中所包含的信息。基于此,本文提出了基于关联词对动态抽取的报道关系检测方法。通过测试得出结论:基于关联词动态抽取的报道关系检测方法非常成功。从实验结果中可以看出,关系约束起着非常重要的作用,本文共采用了 3 个特征:首字母大写特征、位置特征以及距离特征,3 个特征都表现良好。进一步的研究工作包括:定义更加准确地表达报道内容的关系约束,探讨关联词对更加有效的利用方式。

参 考 文 献

[1] Allan J, Lavrenko V, Malin D, et al. Detections, Bounds, and Timelines: Umass and tdt-3[C]// Proceedings of Topic Detection and Tracking (TDT-3). 2000: 167-174

[2] Josef K, Mohamad H, Robert P W. On Combining Classifiers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-239

[3] Chen F, Farahat A, Brants T. Multiple Similarity Measures and Source-pair Information in Story Link Detection[C]// Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, 2004: 313-320

[4] Chen Y J, Chen H H. NLP and IR Approaches to Monolingual and Multilingual Link Detection[C]// Proceedings of the 19th International Conference on Computational Linguistics (COLING2002). Taipei, Taiwan, 2002: 1-7

[5] Brown R D. Dynamic Stopwording for Story Link Detection[C]// Proceedings of Second International Conference on Human Language Technology Research. San Diego, California, 2002: 190-193

[6] Ferret O. Using Collocations for Topic Segmentation and Link Detection[C]// Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, 2002: 260-266

[7] Farahat A, Chen F, Brants T. Optimizing Story Link Detection is not Equivalent to Optimizing New Event Detection[C]// Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL03). 2003: 232-239

[8] Chen F, Farahat A, Brants T. Story Link Detection and New Even Detection are Asymmetic[C]// Proceedings of Human Language Technology Conference (HLT-NAACL 2003). 2003: 13-15

[9] 赵华,赵铁军,张姝,等. 基于内容分析的话题检测研究[J]. 哈尔滨工业大学学报,2006,38(10):1740-1743

[10] 赵华,赵铁军,于浩,等. 面向动态演化的话题检测研究[J]. 高技术通讯,2006,16(12):1230-1235