多文档文摘句子优选算法研究

张 姝　赵铁军　姚 超　郑德权
(哈尔滨工业大学计算机科学与技术学院　哈尔滨　150001)

摘　要：该文通过对文摘句的选择问题进行分析，提出了一种文摘句优选方法，相对于传统的逐个添加句子生成文摘的方法，该文提出的方法是在一定范围内逐个删除句子生成文摘。该方法分两阶段进行句子选择，第 1 阶段获取候选文摘句子集合，采用了直接获取算法和基于冗余信息处理的获取算法。第 2 阶段逐步删除句子，分别以不同特征项作为衡量句子对候选文摘句子集合的贡献，提出了文摘句优选算法。以 DUC2004 为实验语料，通过经句子选择后生成文摘的 ROUGE 得分，验证了句子选择在文摘生成过程中的必要性，与基于冗余信息处理的句子选择方法比较，验证了该文提出算法的有效性。

关键词：句子优选；多文档文摘；冗余信息处理
中图分类号：　TP391　　**文献标识码**：　A　　**文章编号**：1009-5896(2008)12-2921-05

Research on Sentence Optimum Selection Algorithm for Multi-Document Summarization

Zhang Shu　Zhao Tie-jun　Yao Chao　Zheng De-quan
(*Department of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: Analyzing sentences selection in summarization, an approach based on deleting sentences in a sentences set to obtain summary is proposed, which differs from the traditional method of adding sentences to get the summary. It has two stages, one is the process of obtaining the candidate summary sentences set with direct obtaining algorithm and redundancy-based obtaining algorithm, the other is the process of deleting sentences with sentences optimum algorithm. With DUC 2004 as the test corpus, the ROUGE value of summaries gotten by sentences selection proves the necessity of sentences optimum selection for multi-document summarization. Compared with the redundancy-based sentences selection method, the validity of the approach proposed is proved.

Key words: Sentence optimum selection; Multi-document summarization; Redundancy information processing

1　引言

随着互联网的高速发展和广泛普及，网上信息急剧增长，如何对网络信息有效地组织、总结和分析以便快速高效地掌握有价值的信息成为人们日益关注的问题，同时也推动了自然语言处理技术的发展。由于信息量的日益剧增，相关信息、重复信息的频繁出现，单篇文档文摘已经不能减轻用户的阅读量。为向用户提供相关主题的全面扼要的信息，多文档自动文摘技术应运而生。

多文档文摘是这样一个过程，它为了达到一个特定用户或任务的要求，从信息源文本集合中提练出最重要的信息内容以生成一个尺寸缩小的版本。2000 年以后，关于自动文摘技术的相关评测会议和探讨会纷纷出现在自然语言处理领域和信息检索领域知名的会议——ACL，COLING，SIGIR 中[1]。其中由美国国家标准与技术协会(National Institute of Standards and Technology，NIST)支持的文本理解会议[2](Document Understanding Conference，DUC)和日本的搜索引擎评价型国际会议(NIINACSIS Test Collection for IR systems，NTCIR)下的文摘挑战会议[3](Text Summarization Challenge，TSC)主要是致力于为研究者提供一个标准的训练和测试平台，以便对参赛系统进行大规模的评测，使自动文摘技术有更进一步的发展。

在基于句子抽取的多文档文摘方法中比较著名的有 Conroy[4]提出的基于隐马尔可夫模型的多文档文摘，Lin[5]提出的基于单文档结构信息和主题概念特征生成多文档文摘，Blair-Goldensohn[6]提出的句子基于聚类的多文档文摘，以及 Radev[7,8]提出的基于质心的多文档文摘。

目前在多文档文摘的生成过程中主要采用的句子选择方法是按句子权重信息自上而下、逐步筛选与已选文摘句冗余度小的句子添加到文摘中，最终生成符合长度要求的文摘。这种句子选择方法存在以下两个问题：(1)由于文摘长度受限，导致后面权值稍低但包含信息较好的句子没有加入文摘的机会；(2)由于是逐个添加句子构成文摘，所以文摘句的

2007-06-04 收到，2007-11-12 改回
国家自然科学基金(60575041)和国家“863”计划项目(2006AA01Z150)资助课题

选择只是针对当前已选入的文摘句子而非文摘整体的一种判断。由于以上问题，本文提出一种新的文摘句优选方法，该方法首先根据句子权值选出指定文摘长度 n 倍的候选文摘句子集合，然后逐步删除集合中按一定标准对该集合贡献最小的句子，直到剩余的句子长度之和达到指定文摘长度。相对于传统的逐个添加的文摘生成方法，本文提出的方法是在一定范围内逐个删除句子的文摘生成方法。该方法优点在于初选候选文摘句子集合长度的扩大增加了句子加入文摘的概率，而且它的句子删除过程是从候选文摘集合整体考虑的。

为了在 n 倍长度的候选文摘句子集合内加入更多非重复性句子，本文在候选文摘句子集合获取阶段采用了直接获取算法和基于冗余信息处理的获取算法两种算法。在逐步删除句子阶段，分别以不同特征项作为衡量句子对候选文摘句子集合的贡献，提出了文摘句优选算法。在 DUC2004 语料上，实验结果显示了本文提出的句子优选算法的必要性和有效性，分析了算法中参数选取对文摘结果的影响。

本文其他部分的组织结构如下：第 2 节介绍了本文采用的句子加权算法；第 3 节详细阐述了 n 倍长度的候选文摘句子集合获取算法；逐步删除句子阶段的文摘句优选算法；第 4 节是实验与分析；第 5 节是结束语。

2 句子加权模型

在事先分类的文档集合中语义相关的词语可以由同现统计分析获得的前提假设下，Lin 和 Hovy 提出[5]了多文档文摘中主题签名的自动获取。其基本思想是，利用同现统计分析方法从同主题文档集合中自动抽取出与该主题具有强关联的词语集合。统计同现模型采用了对数似然比(Log Likelihood Ratio，LLR)，词语可以是一元、二元或三元的词。经实验表明，经统计分析得到的词语集合是该文档集合主题的很好描述[5]。这种衡量词汇信息权重的方法在多文档文摘中取得了很好的效果。

基于句子抽取文摘方法的基本思想是：找到原文中能够反映中心思想的句子，把这些关键部分抽取出来组织成文本摘要。句子携带信息的重要性一般是根据特征计算来获得的句子权重，特征一般是预先指定的。本文采用了句子位置特征、句子的长度特征和词汇信息特征对句子进行加权计算，加权函数如式(1)所示：

$$\begin{aligned}\mathrm{Weight}(\mathrm{Sent}_i)&=\mathrm{LengthWeight}(\mathrm{Sent}_i)\\&\cdot\mathrm{PositionWeight}(\mathrm{Sent}_i)\cdot\mathrm{SentWordsWeight}(\mathrm{Sent}_i)\end{aligned}\tag{1}$$

$$\mathrm{SentWordsWeight}(\mathrm{Sent}_i)=\frac{\sum_{j=1}^{n_i}\mathrm{WordsWeight}(\omega_j)}{n_i}\tag{2}$$

其中 Sent_i 是文档集合中的第 i 个句子，n_i 为 Sent_i 中包含的词的个数，ω_j 是 Sent_i 包含的单词。位置加权函数 $\mathrm{PositionWeight}(\mathrm{Sent}_i)$ 和句子长度限制函数 $\mathrm{LengthWeight}(\mathrm{Sent}_i)$ 是布尔函数，并根据经验设定只考虑每篇文档的位置为前 10 的句子，和句子长度超过 5 个单词的句子。句子词汇权重函数 $\mathrm{SentWordsWeight}(\mathrm{Sent}_i)$ 是句子包含词汇的权重和与包含词数的比值，其中词汇权重 $\mathrm{WordsWeight}(\omega_j)$ 是采用同现统计分析 LLR 方法计算获得，句子词汇权重函数中除以句子包含词的个数是为了避免长句子因为包含词的个数多而带来的权值偏大的优势。

3 文摘句的选择

在句子加权以后，按照权值将句子降序排列，并对句子进行选择生成文摘。本文对句子的选择分两个阶段完成，第 1 阶段选出指定文摘长度 n 倍的候选文摘句子集合，第 2 阶段逐步删除集合中对该集合新信息贡献最小的句子，直到剩余的句子长度之和达到目标文摘长度。下面分别对第 1 阶段采用的直接获取算法和基于冗余信息处理的获取算法，第 2 阶段对分别以句子所包含的词、主题词以及主题词比率作为衡量句子包含新信息的特征项的文摘句优选算法进行详细阐述。

3.1 候选文摘句子集合获取算法

为生成指定文摘长度 n 倍的候选文摘句子集合，采用了两种获取算法。一种是直接获取算法，即按照句子权值由高到低，依次选取句子加入候选文摘集合中，直到句子长度和满足长度要求。这种算法简单、速度快，但没有对重复或高度相似的句子进行处理，导致句子间冗余信息过多。为了在指定的有限空间内加入更多包含有用信息的句子，引入了基于冗余信息处理的获取算法。该算法是传统的文摘句选取算法－按句子权重从高到低的顺序，选取权值高且与已选入的句子间冗余度低于某阈值的句子。冗余度的定义，直接获取算法和基于冗余信息处理的获取算法将被详细阐述。

本文冗余信息度以候选句子中包含的主题词与已选入句子所包含的主题词的重复度来衡量，其定义如下：

$$\mathrm{Sim}(\mathrm{Sent}_i,S)=\frac{\left|\mathrm{KeyWords}(\mathrm{Sent}_i)\cap\mathrm{KeyWords}(S)\right|}{\left|\mathrm{KeyWords}(\mathrm{Sent}_i)\right|}\tag{3}$$

其中 $\mathrm{KeyWords}(\mathrm{Sent}_i)$ 和 $\mathrm{KeyWords}(S)$ 分别是句子 Sent_i，已选句子集合 S 所包含的主题词的集合，主题词是经第 2 节中词汇权重 $\mathrm{WordsWeight}(\omega_j)$ 计算权重大于 10.83 的词。

算法 1 候选文摘句子集合直接获取算法

输入： 经加权计算后按降序排列的句子序列 A

输出： n 倍指定文摘长度的候选文摘句子集合 S

(1)确定目标长度为 n 倍指定文摘长度；

(2)对于 A 中句子

{

将 A 中句子 Sent_i 作为候选文摘句加入 S 中；

如果 Sent_i 的加入使得 S 中的句子长度和大于目标长度，停止；

}

(3)输出 S；

算法 2　基于冗余信息处理的候选文摘句子集合获取算法

输入：　经加权计算后按降序排列的句子序列 A

输出：　n 倍指定文摘长度的候选文摘句子集合 S

(1)确定目标长度为 n 倍指定文摘长度；

(2)对于 A 中句子

{

如果 $\mathrm{Sim}(\mathrm{Sent}_i, S) < \alpha$，则将 Sent_i 加入 S 中；

如果 S 中的句子长度和大于目标长度，停止；

}

(3)输出 S；

在算法 2 中，α 是冗余度阈值，取值范围在[0，1]之间。当 n 取 1 时，算法 2 是传统的基于冗余信息处理的文摘句选取算法，为了过滤冗余信息，对句子是否加入文摘作判断，这时 α 应该相对较小。当 n 取值大于 1 时，算法 2 作为候选文摘句子集合 S 的选取，只对包含主题词高度相似的句子进行过滤，α 的取值应该较大。

3.2 文摘句优选算法

相对于传统的逐步添加句子构成文摘的句子选择算法，本文采用了在经初步选取的候选文摘句子集合中，逐步删除句子达到指定文摘长度的句子选择算法。句子的删除原则是每次遍历当前的候选文摘句子集合找出对其贡献最小的句子，即删除该句子对集合携带的信息量影响最小。由于在 3.1 节中的候选文摘句子集合选取过程中已经考虑了句子的权重信息，所以文摘句优选算法忽略了句子权重对其的影响，即认为候选文摘句子集合中的句子是同等重要的，这样避免了在指定长度文摘中最后一个文摘句选取时，句子的选择过于依赖句子权重，由于较小的权重差别导致权重稍低但包含有用信息的句子没有加入的资格。由于逐步删除候选文摘句子集合中对其贡献最小的句子，该算法属于整体寻优，而非传统方法对于后续加入句子未知的情况下选择目前最好的句子。

式(4)采用一个句子中包含 S 中其它句子没有的特征项的个数，定义了句子对候选文摘句子集合 S 贡献的衡量，公式如下：

$$\mathrm{Sent}_k = \underset{\mathrm{Sent}_i \in S}{\mathrm{argmin}}\{|\,\mathrm{Token}(\mathrm{Sent}_i) - \mathrm{Token}(\mathrm{Sent}_i) \cap \mathrm{Token}(S \setminus \mathrm{Sent}_i)\,|\} \tag{4}$$

其中 $\mathrm{Token}(\mathrm{Sent}_i)$ 和 $\mathrm{Token}(S \setminus \mathrm{Sent}_i)$ 分别是句子 Sent_i，候选文摘句子集合 S 去除句子 Sent_i 所包含的特征项的集合。这里分别选取了词和主题词作为特征项。

由于文摘长度有限，为了在有限长度内加入更多有用信息的句子，需要考虑句子长度和携带信息量的关系，所以式(5)引入了比率，针对一个句子虽然包含一定量的新特征项，但它同时包含过多的集合已有特征项，导致冗余度过大，句子过长，这样的句子应该在长度和新信息量之间做一个衡量。考虑比率的句子对候选文摘句子集合 S 贡献的衡量公式如下所示：

$$\mathrm{Sent}_k = \underset{\mathrm{Sent}_i \in S}{\mathrm{argmin}} \cdot \left\{\frac{|\,\mathrm{Token}(\mathrm{Sent}_i) - \mathrm{Token}(\mathrm{Sent}_i) \cap \mathrm{Token}(S \setminus \mathrm{Sent}_i)\,|}{|\,\mathrm{Token}(\mathrm{Sent}_i)\,|}\right\} \tag{5}$$

算法 3　文摘句优选算法

输入：　n 倍指定文摘长度的候选文摘句子集合 S

输出：　指定文摘长度的文摘句子集合 S_Target

(1)获取 S 中特征项集合 W；

(2)循环

{

遍历 S 中句子，寻找对集合贡献最小的句子 Sent_k，记录该句子；

更新长度集合句子长度和；

如果长度小于指定文摘长度，则停止；

否则，更新句子集 S——删除 Sent_k；更新 S 中特征项集 W——删除 Sent_k 包含的新的特征项；

}

(3)输出 S 为指定文摘长度限制下的文摘句子集合 S_Target；

4　实验与分析

4.1 实验语料与评测机制

本文采用 DUC2004 中任务 2，即短的多文档文摘语料进行实验研究，语料情况如下：

语料主要来自话题检测与跟踪(Topic Detection and Tracking，TDT)任务，其中的文档来源于美联社新闻和纽约时报。测试语料共有 50 个英文文档集合，每个集合包含大约 10 篇文档和 4 篇人工文摘。其中人工文摘由 NIST 评价人员根据文摘长度要求的限制生成，且文摘仅来源于所在文档集合中的内容。

任务 2 要求参赛系统为每个文档集合生成指定长度的文摘，长度要求为 665byte。并规定，对于超过该长度的文摘在评价过程中自动剪切至 665byte，对于没有达到长度要求的文摘也没有任何奖励。提交的系统文摘采用 ROUGE 进行自动打分。

在 DUC2004 中，一种自动评价方法 ROUGE 被采用并作为主要的评价方法[9]。ROUGE-N 是一种基于召回率的 n-gram 同现统计方法。ROUGE-N 计算系统生成文摘与人工文摘之间同现单元的个数。计算公式如下：

$$\text{ROUGE-}N = \frac{\sum_{S \in \{\mathrm{referenceSummaries}\}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)}{\sum_{S \in \{\mathrm{referenceSummaries}\}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}(\mathrm{gram}_n)} \tag{6}$$

其中 n 是 gram_n 中 n-gram 的元，$\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)$ 是系统文摘与人工文摘之间共现的 n-grams 的个数。这里选取了

1-gram 和 2-gram 的 ROUGE 得分作为文摘的评分。在 DUC2004 中，所有文摘方法的 ROUGE-1 得分在[0.24190，0.38224]之间，ROUGE-2 得分在[0.01876，0.09216]之间。

4.2 实验结果与分析

为了验证本文提出的文摘句优选算法的必要性和有效性，设计了以下两组实验，其中所有实验的前期句子加权方法都采用第 2 节的句子加权模型，以经句子选择后得到的文摘的 ROUGE 得分来衡量句子选择方法的性能。

第 1 组实验，验证基于本文提出的文摘句优选算法(算法 1+算法 3)的必要性和有效性，Baseline-1 句子选取过程为将经过权重计算的句子降序排列、由前往后依次加入句子直到文摘指定长度。结果如表 1 所示，其中参数 n 是算法 1 中的 n 倍指定文摘长度，这里分别选取了 2–5 倍长度进行实验，在算法 2 中贡献最小句子的衡量式(4)和式(5)的特征项分别选取了词、主题词和主题词比率。

表 1 基于算法 1 加算法 3 的文摘 ROUGE 得分

参数 n	特征项	ROUGE-1	ROUGE-2
n=2	词	0.36844	0.08692
	主题词	0.37441	0.09319
	主题词比率	0.37819	0.09427
n=3	词	0.37527	0.08665
	主题词	0.38914	0.09673
	主题词比率	0.38737	0.09297
n=4	词	0.37184	0.08419
	主题词	0.39019	0.09648
	主题词比率	0.38510	0.09029
n=5	词	0.36680	0.08026
	主题词	0.37886	0.08659
	主题词比率	0.39057	0.08766
Baseline-1	—	0.35217	0.08182

由表 1 可以看出：

(1)本文提出的文摘句优选方法在所有的参数设置下的得到的文摘的 ROUGE-1 和 ROUGE-2 得分都比 baseline-1 高，甚至多个 ROUGE-1 得分超过 DUC2004 中的最高得分，这说明在相同的前期文档分析句子加权后，文摘句子优选的必要性以及本文提出文摘句优选方法的有效性。

(2)对于 n 的选取。当 n 取 3 和 4 时候，不同特征项的 ROUGE 得分都比较高。由于没有考虑去除重复的句子，所以此时在候选文摘句子集合 S 中存在一定的重复句子，这使得 n 取值较大才能包含更多的有用信息句子。

(3)对于特征项的选取。整体来看，以词作为特征项的效果相对不好，因为句子中含有的词未必是对文档集合有意义的词，所以以它来衡量句子包含新的重要信息效果不好。而以主题词和主题词比率为特征项，ROUGE 得分相当。

第 2 组实验，验证基于本文提出的文摘句优选方法(算法 2+算法 3)的有效性，Baseline-2 方法为传统的基于冗余信息处理的文摘句选择方法(算法 2)。实验结果如表 2，表 3 所示，表 2 和表 3 分别对应于 2 和 3 倍指定文摘长度生成候选文摘句子集合，Baseline-2 方法则是直接生成指定文摘长度，为

表 2 基于算法 2 和算法 3 的文摘 ROUGE 得分(n=2)

参数 α	特征项	ROUGE-1	ROUGE-2
0.9	Baseline-2	0.37215	0.09080
	词	0.37638	0.09167
	主题词	0.38518	0.09382
	主题词比率	0.39286	0.09605
0.8	Baseline-2	0.37213	0.09140
	词	0.37795	0.09228
	主题词	0.38285	0.09272
	主题词比率	0.38753	0.09238
0.7	Baseline-2	0.37450	0.09211
	词	0.37232	0.08692
	主题词	0.37940	0.09223
	主题词比率	0.38588	0.09233
0.5	Baseline-2	0.38447	0.09309
	词	0.36398	0.08121
	主题词	0.37123	0.08411
	主题词比率	0.36572	0.07931

表 3 基于算法 2 和算法 3 的文摘 ROUGE 得分(n=3)

参数 α	特征项	ROUGE-1	ROUGE-2
0.9	Baseline-2	0.37215	0.09080
	词	0.36860	0.08260
	主题词	0.38377	0.09304
	主题词比率	0.38953	0.09014
0.8	Baseline-2	0.37213	0.09140
	词	0.37080	0.08415
	主题词	0.38901	0.09401
	主题词比率	0.39189	0.08768
0.7	Baseline-2	0.37450	0.09211
	词	0.36061	0.07416
	主题词	0.38078	0.08710
	主题词比率	0.38179	0.08319
0.5	Baseline-2	0.38447	0.09309
	词	0.34108	0.06420
	主题词	0.35741	0.07597
	主题词比率	0.32596	0.05530

便于比较，在两个表中重复给出。对于算法 2 中的冗余度阈值 α 分别取了 0.9，0.8，0.7 和 0.5。

结合表 2 和表 3 可以看出：

(1)以词为特征项的结果没有以主题词和主题词比率为特征项的结果好，再次说明了词不适合作为特征项来衡量句子包含新的重要信息。

(2)冗余度阈值 α 选取对结果影响较大，当 α 取 0.8 和 0.9 时，以主题词和主题词比率为特征项的结果明显好于 Baseline-2，而且算法 2 的引入，使得该结果与表 1 中的对应项相比也有所提高。说明了相对于算法 1 直接获取候选文摘句子集合，算法 2 的引入，去除高重复句子，在一定程度上扩大候选文摘句子集合，改善了文摘结果。

(3)当冗余度阈值 α 取值为 0.5 时，Baseline-2 方法效果最好，这是由于候选文摘句子集合已经经过大幅度筛选，句子间的冗余度较小，且句子间权值差异较大，而本算法在逐步删除句子时没再考虑句子权值，所以导致权值小的句子加入，使得文摘结果下降。

(4)Baseline-2 方法最高 ROUGE-1 得分为 0.38447，而本文提出的句子优选方法得分大于此分数的结果很多，所以说明了本文提出方法的优越性。

5　结束语

本文对文摘句的选择问题进行研究，提出了一种文摘句优选方法，在一定句子范围内逐个删除句子的最终生成文摘方法。该方法首先根据句子的权值选出指定文摘长度 n 倍的候选文摘句子集合，然后按一定规则逐步删除对集合贡献最小的句子，直到剩余的句子长度之和达到指定文摘长度。通过在 DUC2004 语料上的两组实验，经过句子优选得到的文摘 ROUGE 得分比简单的根据句子权重选择文摘(Baseline-1)有明显提高，验证了句子优选的必要性。与基于冗余信息处理的句子选择方法(Baseline-2)比较，证明了本文提出算法的有效性。同时分析了文摘句优选算法中不同特征项的选取对文摘 ROUGE 得分的影响，得出单纯的词汇信息不适合作为特征项来衡量句子对候选文摘句子集合新信息贡献的结论。

参 考 文 献

[1] Lin C Y and Hovy E. From single to multi-document summarization: A prototype system and its evaluation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 2002: 457–464.

[2] DUC. Document Understanding Conferences. http://duc.nist.gov.

[3] NTCIR. NII Test Collection for IR Systems. http://research.nii.ac.jp/ntcir/.

[4] Conroy J M and Schlesinger J D. Left-brain/right-brain multi-document summarization. Proceedings of the 2004 Document Understanding Conference (DUC 2004), Boston, MA, May 6-7, 2004.

[5] Lin C Y and Hovy E. The automated acquisition of topic signatures for text summarization. Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000), Saarbrücken, Germany, July 31–August 4, 2000: 495–501.

[6] Blair-Goldensohn S and Evans D, *et al*.. Columbia university at DUC 2004. Proceedings of the 2004 Document Understanding Conference (DUC 2004), Boston, MA, May 6-7, 2004.

[7] Erkan G and Radev G R. The university of Michigan at DUC 2004. Proceedings of the 2004 Document Understanding Conference (DUC 2004), Boston, MA, May 6-7, 2004.

[8] Otterbacher J C, Winkel A J, and Radev G R. The Michigan single and multi-document summarizer for DUC 2002. Proceedings of the 2002 Document Understanding Conference (DUC 2002), Philadelphia, Pennsylvania, July 11-12, 2002.

[9] Lin C Y. ROUGE: A package for automatic evaluation of summaries. Proceedings of the ACL 2004 Workshop on Text Summarization, Spain, 2004: 74–81.

张　姝：　女，1977 年生，博士，主要研究方向为自动文摘.

赵铁军：　男，1962 年生，教授，博士生导师，主要研究方向为自然语言处理、机器翻译.

姚　超：　男，1981 年生，硕士生，研究方向为自动文摘.

郑德权：　男，1968 年生，副教授，主要研究方向为信息检索、知识获取.