# 基于重要句群检索性能比较研究

何　琳　黄水清　徐彩琴

**摘　要**　重要句群是指最能表达文献主题的若干句子的集合,客观性强、生成效率高,可在利用自动文摘成果的基础上基于重要句群进行检索。在对句子进行预处理、文献语词权重计算以及句子权重计算后生成重要句群。利用基于向量模型的方法以及构建检索式和检索提问分别对基于文摘、重要句群和全文的检索性能进行对比分析。实验结果表明,基于重要句群的检索性能总体优于作者文摘,但是低于全文,可以将重要句群作为后台数据或搜索引擎的摘要等。句子级别的文本处理对提升文本检索效果的作用不大,而把文本处理提升至上下文的语义级别是可能的有效方法。图1。表4。参考文献9。

**关键词**　重要句群　检索性能　检索评价　文献检索　比较研究
**分类号**　G354.2

**ABSTRACT**　Important sentence groups are sets of several sentences. It is objective with a high generating efficiency. We can retrieval based on them using results of automatic abstraction. Important sentence groups can be generated after pretreating sentences, weighting literature words and sentences. The paper compares and analysizes abstracts' retrieval effectiveness with sentence groups and full text making use of VSM and constructing search strategy and questions. It concludes the result that the effectiveness of sentence group is better than abstract, worse than full text, which can be used as backstage database or abstract of search engine. However, text processing of sentences level doesn't play a significant role on promoting text retrieval effectiveness. Paragraph retrieval may be a useful approach to improve the effective way. 1 fig. 4 tabs. 9 refs.

**KEY WORDS**　Important sentence groups.　Retrieval effectiveness.　Retrieval evaluation.　Document retrieval.　Comparative study.
**CLASS NUMBER**　G354.2

## 1　引言

与其他的基于文献内容的检索途径相比,基于文摘检索除了方便灵活外,还具有更高的查全率和查准率,以及更短响应时间等优势。国内的期刊检索系统,如中国期刊网、维普资讯、万方数字化期刊数据库等都提供了基于文摘检索的途径。然而,并不是所有的文献都带有文摘,且文摘带有一定主观性和片面性,质量也不同。是否能够从文献中抽取重要句子组成重要句群代替文摘进行检索?重要句群的检索效率是否优于文摘检索?本文提出一种重要句群的生成方法,并选取不同学科的文献进行重要句群、文摘和全文检索的检索效率比较研究,并给出结论。

## 2　基于重要句群检索的可行性

重要句群是指最能表达文献主题的若干句子的集合[1]。重要句群与文摘相比,都具有集中表达文献主题的特点,不同之处在于文摘是由作者或文摘员提炼而成,具有很好的可读性和连贯性,同时也不可避免地带有一定的主观性和片面性;重要句群是由若干重要句子组成,可读性和连贯性较差,但不带有任何主观色彩,且生成效率高。

计算机基于文摘的检索优势是由于文摘能够较好地揭示文献的主题,在这一点上重要句群与文摘类似。此外,信息检索用户的真正目的是通过文摘检索来最终获取原始文献,文摘只是计算机检索的中间过程,因此用户并不是

特别关心文摘的可读性与否。近年来,自动文摘技术成为国内外信息检索领域的研究热点,基本能够较准确抽取文献的主题,主要不足在于自动文摘的可读性和连贯性差[2]。而重要句群不需要考虑文摘的结构、语义、语用以及连贯性和可读性问题,可以很好地利用自动文摘取得的技术成果,因而基于重要句群检索在技术上是可行的。

## 3 重要句群的提取方法

重要句群提取的基本方法是对句子进行重要性统计。首先对文献进行分词,计算文献语词权重;其次是提取句子相关信息,包括句子长度、位置、是否有提示词等;然后综合以上信息计算句子权重,将句子根据权重进行重要性排序,根据文摘的长度决定最后生成的重要句群的长度。

### 3.1 预处理

预处理主要包括对文献格式和内容的处理以及分词词典的构建。通过预处理可以帮助计算机识别数据,同时分析文献集合的写作特点。

### 3.2 文献语词权重计算

文献语词计算主要是以篇为单位对待处理文献集合进行词频统计。首先选定具有表达能力的文献语词,再根据 TF-IDF(term frequency-inverse document frequency)方法计算文献语词的权重。为了提高语词的质量,本文选取了词长在 2 字词以上的语词,计算公式如下:

$$W(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_t+0.01)]^2}}$$

其中,$W(t,d)$为词 $t$ 在文献 $d$ 中的权重;$tf(t,d)$为词 $t$ 在文献 $d$ 中的词频;$N$ 为训练文献的总数;$n_t$ 为训练文献集中出现 $t$ 的文本数;分母为归一化因子。

### 3.3 句子权重计算

句子权重计算是重要句群提取的重要步骤。句子权重的计算受句子长度、位置信息以及提示词和废弃指示词的影响。

(1)句子位置信息

不同位置的句子揭示文献主题的能力是不同的,Salton[3]认为文章的中心段落可以作为文摘的核心;巴森代尔[4]经过大量的统计发现论题句作为段首句出现的占 85%,以段末句出现的占 7%,因此中心段落的段首句应该占有更高的权重。经过对科技文献的分析发现,总起段和总结段为科技文献的中心段落。

(2)提示词

在科技文献中,常常会有许多短语或词汇引申出文献主题的总结性语词,称为提示词[5],如"本文论述了"、"综上所述"、"研究结果表明"等等。含有这些提示词的句子都能够更好地反映文献的主题内容,在计算句子权重的时候应该提高它们的权值。

(3)废弃指示词

与提示词相对应的是废弃指示词,如"例如"、"举例来说"等,含有这些词的句子在表达确切文献主题作用上就要大打折扣,因此在计算句子权重的时候,要降低它们的权值。

通过对以上指标的分析,给出文本句子权重的计算公式:

$$W(S)=\mu_p\mu_s\frac{\sum_{i=1}^{n}W_{is}}{l_s}$$

其中,W(S)表示句子的权值;$W_{is}$表示句子 S 中 i 的权值;ls 表示句子的长度;$\mu_p$ 表示句子的位置信息,如果句子位于第一段取值 1.4,最后一段取值 1.2,其余情况取值 1;$\mu_s$ 表示句子还有指示词的情况,如果还有指示词取值 2,如果还有废弃指示词取值 0.5,其他情况取值 1。

### 3.4 重要句群的生成

按照句子的权值对句子进行重要性排序。文本越长含有的信息量就可能越大,被检索到的可能性就越大。为了与文献进行对比研究,本文在重要句群生成的时候,根据文摘长度的大小进行截取,重要句群的长度与文摘的长度差值控制在 20 个汉字以内,如果超过长度将对重要句群中的句子进行强制截断。

## 4 测试分析

### 4.1 测试方法

4.1.1 内部测试

内部测试方法是对生成的重要句群与全文进行相似度比较,为了客观衡量重要句群的质量,把文摘与全文的相似度作为对比数据。测试的方法是采用基于向量空间模型的文本相似度计算方法,分别将全文、文摘、重要句群看作是一组由正交词条所组成的矢量空间,对三者进行分词、特征抽取、权值计算,通过余弦夹角来计算相似度。夹角越小说明相似度值越高,与全文表达的语义越接近,其质量也越高。

4.1.2 外部测试

为了克服内部评价方法的单一,同时采用外部测试的方法,即基于特定检索提问的检索性能比较。测试的基本方法是在相同的检索提问条件下,分别将全文、重要句群、文摘作为测试集,以定量的方式测试在相同检索提问下三者所表现出的检索性能。检索性能的测评指标采用传统的查全率、召回率和 F1 值。

### 4.2 数据来源

4.2.1 测试集

为了准确、全面和客观地对基于重要句群的检索性能进行研究,测试集的构建选取了三个学科门类。其中,理学选择了畜牧兽医,管理学选择了经济学,工学选择了计算机科学,选择文献共计 2774 篇,分别对由不同学科构成的文摘、全文和重要句群的检索性能进行对比分析。

4.2.2 检索提问及相关性判断

为了体现研究的准确性和客观性,检索提问选取了高校图书馆 2002 – 2006 年的真实查新课题以及 2002 – 2007 年国家社会科学、自然科学基金相关项目名称共同构成检索测试的 57 个检索提问,年份分布均匀,具有代表性,客观性较高。

相关性判断一直以来都是信息检索领域有争议的问题。为了避免个人因素的影响和理解的偏差,提高文献相关性判断的准确度,测试中文献的相关性判断选取三位与本研究无关的专业人员进行文献相关性判断,将三位专业人员的判断结果进行汇总修正,以得到尽可能准确、客观的判断结果。

### 4.3 测试结果

4.3.1 内部测试结果

表 1 展示了基于向量空间模型(Vector Space Model,VSM)的相似度的比较值,从表中数据来看,重要句群与全文的平均相似度为 0.54,高于文摘与全文的平均相似度 0.49,证明重要句群的总体质量高于文摘。从不同学科来看,经济类、畜牧类文献的重要句群的质量明显高于文摘的质量,而计算机类文献的重要句群质量低于文摘的质量,这是因为计算机类文献的文摘通常比较详细,而且会在文摘中将文章各部分内容作提要,因此文摘质量比较高。即便如此,虽然重要句群相似度值低于文摘与全文的相似度,但差别并不是太大,从这一点也可以说明,通过内部测试方法得到的重要句群的质量相对比较高。

**表 1 各学科文摘、重要句群分别与全文的向量相似度对比**

| 测试学科 | 测试项 | 0 – 0.3 | 0.3 – 0.5 | 0.5 – 0.7 | 0.7 – 1 | 相似度均值 |
|---|---|---|---|---|---|---|
| 畜牧兽医 | 文摘 | 0.03 | 0.24 | 0.45 | 0.28 | 0.59 |
| | 句群 | 0.01 | 0.07 | 0.41 | 0.51 | 0.69 |
| 经济学 | 文摘 | 0.84 | 0.13 | 0.03 | 0 | 0.28 |
| | 句群 | 0.19 | 0.53 | 0.27 | 0.01 | 0.36 |
| 计算机科学 | 文摘 | 0.03 | 0.18 | 0.45 | 0.35 | 0.62 |
| | 句群 | 0.07 | 0.25 | 0.42 | 0.25 | 0.58 |

4.3.2　外部测试结果

分别将文摘、重要句群以及全文作为检索来源,由事先选定的 57 个检索提问对其进行检索,将各自的检索结果按照查准率、召回率和 F1 值三个指标进行计算,结果如图 1 所示。通过图 1 看出总体检索性能全文最优,重要句群次之,文摘最差。

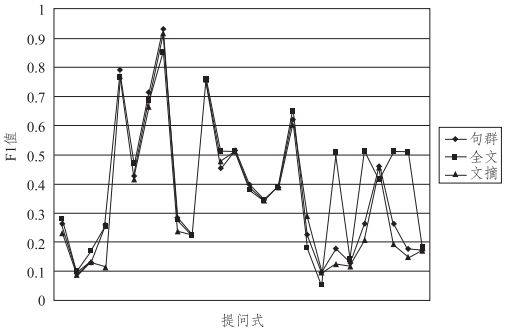


**图 1　重要句群、文摘、全文 F1 值对比**

表 2 为重要句群、全文、文摘三者对于相同的检索提问所得到的检索评价的平均值。

**表 2　文摘、句群、全文的总体 P、R、F1 平均值对比**

| 测试项 | Precision | Recall | F1 |
|---|---|---|---|
| 文摘 | 0.505 | 0.353 | 0.429 |
| 句群 | 0.551 | 0.507 | 0.529 |
| 全文 | 0.558 | 0.654 | 0.606 |

为了准确衡量不同学科文摘和重要句群的质量,表 3 统计了文摘、重要句群和全文在不同学科中的检索结果。统计结果表明:经济类文摘的检索性能最差,重要句群在这个学科中的检索性能远高于文摘,低于全文;畜牧类中,重要句群的检索性能最高,优于文摘,略高于全文;计算机类的文摘质量相对比较高,因而表现出了不错的检索性能,计算机类中重要句群的检索性能略低于文摘的检索性能。

**表 3　各学科文摘、句群、全文的 P、R、F1 平均值对比**

| 测试学科 | 测试项 | Precision | Recall | F1 |
|---|---|---|---|---|
| 畜牧兽医 | 文摘 | 0.475 | 0.399 | 0.437 |
| | 句群 | 0.625 | 0.522 | 0.573 |
| | 全文 | 0.5 | 0.550 | 0.525 |

续表

| 测试学科 | 测试项 | Precision | Recall | F1 |
|---|---|---|---|---|
| 经济学 | 文摘 | 0.575 | 0.175 | 0.375 |
| | 句群 | 0.475 | 0.560 | 0.518 |
| | 全文 | 0.275 | 0.945 | 0.610 |
| 计算机科学 | 文摘 | 0.565 | 0.484 | 0.525 |
| | 句群 | 0.555 | 0.438 | 0.496 |
| | 全文 | 0.8 | 0.466 | 0.633 |

## 4.4　结论

4.4.1　重要句群的总体检索性能优于文摘的检索性能

从表 1 看出,重要句群与全文的相似度高于人工文摘与全文的相似度。重要句群的句子完全来自文献正文,在重要句群的抽取中包含了文献的重要结论或创新点,较为客观地反映了文档的重要内容。而人工文摘带有一定的主观性,仅仅重复文章中的某个段落,对于论文的价值、重要性、结论或创新点总结不足。以经济学为代表的社会科学类的期刊论文的文摘最为典型,非常简短、概括,检索后往往得不到任何实质性的信息,降低了文摘检索的作用。通过测试不难发现,重要句群的检索性能在某些学科远高于人工文摘,虽然某些学科略低,但至少与文摘检索性能基本相差不大。

因此,通过本测试可以得出,在期刊论文数据库建设中,特别是对于文摘质量较差或者缺乏作者文摘的论文完全可采用重要句群代替人工文摘进行检索的方式,以提供给检索者更多指引检索的信息。另外,目前的搜索引擎只是简单截取文档的前几行或者将含有查询关键词的语句抽取出来作为摘要返回给用户,显示结果也多为用户感兴趣的内容,但很多情况下摘取出来的句子表达的并不是该网页的中心内容。而重要句群基本上客观准确地反映了全文的主要内容,且抽取重要句群的花费少,速度快。如果将基于用户查询的重要句群的检索结果应用于搜索引擎摘要的显示将会对用户有更实质性的指引,从而进一步提升搜索引擎的信息服务质量。

4.4.2　基于语义级的处理是提升全文检索的

关键

全文检索虽然总体性能高于重要句群和人工文摘,但相对于重要句群检索而言,全文的存储需要占据很大的空间,检索需要更长的响应时间。特别是现在每天都有海量数据产生,数据库的数据规模庞大,检索系统的构建要综合考虑检索精度、检索速度以及检索成本,相比之下,重要句群可把文献的词数压缩到10%—20%。重要句群虽然比全文的检索性能略低,但检索响应时间与存储空间都会大大降低。考虑到检索的效率和检索精度,基于重要句群的检索如果在某种程度上可以做一定的改进,则在某些情况下不失为一种全文替代品。

重要句群的检索性能低于全文的主要原因在于目前的信息检索机制是建立在"bag of words"的基础上,停留在字词匹配的层面。本文生成重要句群的主要目的是为了与人工文摘进行比较,所以其长度是参照文摘的长度。在信息检索中,检索效果是与文本的长度相关的,重要句群的长度远远低于全文的长度,包含的信息量(关键词)远远少于全文,这是重要句群检索性能低于全文的一个原因。

研究发现,重要句群标引全文的索引强度取决于文档本身的规模。文档规模小,包含的主题相对单一,重要句群表达全文的能力就好;反之,就会降低。产生这种问题的原因在于重要句群生成时是以文档实体为一个抽取单元,将全文处理成为若干个句子,无论文档包含多少主题,句群的生成都是将文档作为一个完整的单元进行处理和描述。从这个角度看,重要句群相对于全文而言,仅是对全文的删节,将文档浓缩的同时,也丢失了部分信息。

从本文得到的启示是,信息抽取的对象停留在文档级别,对检索效果的提升作用不大。如果能对文档进行适当的语义划分,将处理单元缩小为段落或一定的主题块,就会使检索效果提升。国外有研究证明,"全文+段落检索"的方式对检索效果的提升有一定的帮助[7]。段落检索结合全文或文摘的检索性能优于单一的全文或者文摘检索,而段落检索结合全文的效果在几种检索方式中是最突出的,如表4所示。国内也有学者研究基于篇章结构的自动文摘的提取[8-9],将文本的处理从句子级提升到基于主题的上下文级别。

目前全文检索在发展中还存在着一定的不足,将全文作为整个的标引单元进行处理,从实验来看对检索的帮助不大,因此对全文做一定的语义单元的拆分(例如段落检索)是解决检索困难的可能途径之一。这也符合目前用户对检索内容的部分内容感兴趣的要求。

**表4 文摘、全文和段落检索的平均查准率比较[7]**

| 项目 | Ivory(bm25) | Ivory(Luncene) |
|---|---|---|
| 文摘 | 0.163 | 0.129 |
| 全文 | 0.146 | 0.235 |
| 段落 | 2.240 | 0.206 |
| 段落+文摘 | 0.257 | 0.216 |
| 段落+全文 | 0.257 | 0.262 |

## 5 结语

本文在借鉴自动文摘成果的基础上,从文本中抽取重要句群,并对重要句群的检索性能同人工文摘和全文进行了对比实验。实验证明,基于重要句群的检索性能优于人工文摘,可以考虑将重要句群作为后台数据或搜索引擎的摘要等,能够得到更准确的检索结果。另外,通过对重要句群和全文的检索性能的对比实验发现,仅对全文进行句子级别的处理,对提升检索效果帮助不大,对文本进行语义级别的处理,如结合段落检索是可能的有效方式。需要指出的是,研究的出发点是为了将重要句群、文摘和全文的检索性能作对比,因此在检索时所采用的向量空间模型没有进行优化,得到的检索结果相对较低。此外,本文所采用的重要句群的生成方法没有考虑到同义词等因素的影响,在下一步的工作中,可以对重要句群的抽取方法进行优化,进一步提升重要句群的生成质量,同时可探讨融合段落检索的文本检索的检索效果。

**参考文献:**

[1] Ren F. J. Automatic abstracting important sentences[J]. International Journal of Information Technology & Decision Making, 2005, 4(1): 141-152.

[2] 黄丽琼. 中文自动文摘及评价方法的研究[D]. 重庆: 重庆大学, 2007.

[3] [美]哈罗德·博科, 查尔斯. L. 贝尼埃. 文摘的概念与方法[M]. 赖茂生, 王知津, 译. 北京: 书目文献出版社. 1991:4-5.

[4] Baxendale, P. E. Machine-made index for technical literature—an experiment[J]. IBM. Journal of Research and Development, 1958, 2(4): 354-361.

[5] 苏新宁. 信息检索理论与技术[M]. 北京: 科技文献出版社, 2004: 311-316.

[6] 徐彩琴. 基于重要句群与基于作者文摘的汉语文献检索比较研究[D]. 南京:南京农业大学, 2008.

[7] Jimmy Lin. Is searching full text more effective than searching abstracts? [EB/OL]. [2009-02-15]. http://www. biomedcentral. com/1471-2105/10/46.

[8] 王建波,杜春玲. 基于篇章理解的自动文摘研究[J]. 中文信息学报, 1995(3):33-42.

[9] 杨建林. 一种使用自动聚类思想的自动文摘方法[J]. 情报学报, 2001(5):532-536.

**何 琳** 南京农业大学信息管理系讲师。通讯地址:南京市童卫路6号。邮编 210095。

**黄水清** 南京农业大学信息管理系教授,硕士生导师。通讯地址同上。

**徐彩琴** 南京农业大学信息管理系。通讯地址同上。

(收稿日期:2009-01-09;修回日期:2009-03-03)

# "2009 搜索行为与用户认知研究北京研讨会"成功举办

2009 年 6 月 27 日,"2009 搜索行为与用户认知研究北京研讨会"在北京大学成功举办。会议由北京大学信息管理系暨国家信息资源管理北京研究基地和南京理工大学经济管理学院信息管理系共同主办,国家信息资源管理北京研究基地承办,来自北京大学、南京理工大学、中国科技信息研究所、中国人民大学、北京师范大学、南京大学、西南大学、武汉大学、西安电子科技大学的专家和研究生,《中国图书馆学报》和《情报学报》等专业期刊的代表,以及互联网企业的代表共 70 余人参加。

搜索行为和用户认知研究是信息管理和情报学中一个非常重要的研究领域。我国在这方面虽起步较晚,但近年来发展较快,一些高校、企业和研究院所先后建立了专门的研究机构或团队,并陆续推出了一批研究成果。此次研讨会是国内信息管理与图书情报学界首次以搜索行为和用户认知研究为主题的学术研讨会。会上,南京理工大学经济管理学院研究团队报告了该团队在分类理解、隐喻认知、心智模型、强化学习模型等方面的研究成果,北京大学信息管理系研究团队报告了其近 4 年来在语言使用行为、日志挖掘和社会网络方面的研究成果和进展。来自中国科技信息研究所、中国人民大学、西南大学和北京师范大学的专家分别介绍了他们的研究情况。艾瑞、谷歌、百度和中国制造网等互联网企业的代表也报告了各自公司的最佳实践。与会专家和代表们就各自所关心的问题进行了热烈的讨论。

(北京大学信息管理系)