偏重摘要及其实现技术*

刘功申[1] 胡佩华[1] 岳奕[2] 王永成[1]

[1]（上海交通大学计算机科学技术系，上海，200030）；
[2]（华东师范大学信息科学系，上海，200062）
E-mail: lgshen@sjtu.edu.cn

摘要：偏重摘要系统是一个非常有意义的研究课题。本文实现了用于概念扩展的概念库，并以此为基础提出了偏重摘要系统的实现方法。在偏重摘要的生成过程中，讨论了主题相关加权和偏重相关加权，并且通过一定的策略选取句子形成最终摘要。实验证明，系统对绝大部分文章都能做出既满足用户偏重要求，又能在一定程度上反映全文主题的摘要。

关键词：概念扩展，偏重摘要，自然语言处理

1 引言

传统意义上的摘要是对全文信息的浓缩，是对原文所描述的主题、范围和结果的一种简洁概括。这种摘要是静态的，不能适应用户的个性化要求。当前，国内外的大多数自动摘要系统所产生的摘要都是传统意义上的摘要。与传统的摘要比较，偏重摘要不仅仅决定于原文的主题，而且也决定于用户的个性化要求。对一篇特别长的文章，如果用户只关心某一方面的问题（例如工业），这就涉及到偏重问题。偏重问题是实现用户个性化摘要必不可少的技术。

实现偏重摘要具有两个现实意义：第一，在形成偏重摘要的过程中，强调用户的要求，使摘要结果能满足用户特殊要求；第二，在搜索引擎系统中，可以根据查询要求返回一个简短的摘要。用户可以快速浏览这个简短的结果，来判断文档与查询要求的相关度[7]。在判断相关度方面，由于偏重摘要考虑了原文主题和用户的查询两个方面，偏重摘要比传统摘要和现有搜索引擎所提供的方式更加可靠。

从1958年，H.P.Luhn[1]在IBM704机器上进行第一次自动摘要至今，自动摘要已取得很大的发展。从80年代末，我所在的实验室就开始了自动摘要系统的研究工作，并取得了“世界领先水平”。该课题的是在原有系统的基础上进行的，很多地方利用了原有的成果。

本文首先介绍了概念库的结构和建造方法，利用该概念库可以将用户输入的个性要求转换成用户所期望的相关概念。然后，根据这些概念来对文章进行偏重摘要。在偏重摘要的生成过程中，讨论了主题相关加权和偏重相关加权两个部分，并且通过一定的策略选取句子形成最终摘要。

在接下来的一个部分，本文介绍了偏重摘要系统的系统结构。概念库的组织方式及其访问算法在第3部分描述。在第4部分，本文讨论了句子的主题相关加权、偏重相关加权和句子选取策略等。第5、6部分是相关试验、结论以及前景展望。

*本课题受到国家自然科学基金资助(60082003)。

2 系统结构

与一般的摘要系统不同，偏重摘要系统需要用户的输入信息。在产生摘要过程中，这种方法能够把焦点放在用户关心的部分，而不是把原文的各个部分平等对待。因此，偏重摘要系统需要接受用户输入（在搜索引擎中，把查询关键字串当作用户输入）来获取用户关心的焦点。为了避免用户输入的片面性和局限性，对用户输入进行概念扩展是必须的。例如，有一篇很长的政府工作报告，报告中涉及工业、农业、医药卫生等各个方面。用户（可能是分管工业的同志）可能只关心有关工业的论述，系统可以为用户产生一个聚焦工业方面的摘要。如图 1 所示，偏重摘要系统的流程如下：

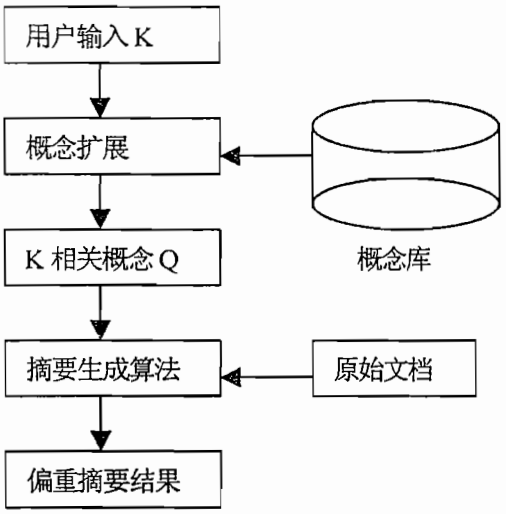

图 1. 偏重摘要系统结构图

- 输入：接受用户输入集合 K，K 是用户关心的关键字串集合。
- 概念扩展：根据概念库，对用户输入的关键字串集合进行概念扩展。该模块避免了用户输入的片面性和局限性。例如，用户输入“计算机”，系统则认为“电脑”等都是相关概念。Q={计算机，电脑，……}。
- 摘要生成算法：首先，该模块根据一定的策略计算原文各个句子的分值。在计算分值时，考虑了全文因素和用户输入因素两个方面。然后，选取分值较高的句子形成摘要候选句集合。最后，对候选句集合进行平滑处理，最终形成摘要结果。

3 概念库

偏重摘要系统把用户输入作为用户关心的焦点。由于用户输入的片面性和局限性以及相关信息在原文中表达形式的多样性，相关信息在原文中的表达形式不可能被用户输入概括。为了克服这种问题，概念扩展是必须的环节。HowNet、WordNet 和同义词词林都在不同程度上揭示和实现了概念之间的关系，这些都是建设概念库的基础。但是，由于建库的目的和侧重点不同，这些技术都不可能满足本文所要求的概念扩展功能。

定义 1(概念条目)：某个代表字串及其各级扩展字串组成的字串集合叫做一个概念条目。概念库是由各个概念条目按一定的组织方式组成的集合。

在深入研究 HowNet、WordNet 和同义词词林的基础上，本文借鉴了它们的优点和长处，提出了适用于偏重摘要的概念库组织方式。表 1 是对各扩展层次的定义和描述。其中，六级扩展是一个虚拟节点，实

际上，它并没有存储在概念库中。在应用过程中，如果需要第六级扩展，可用当前概念条目中的字串递归访问概念库获得。从表 1 可以看出，概念条目和零、一、二、三级扩展字串有较高同义度；和四、五级扩展字串的同义度较低；和六级扩展字串的同义度最低。一个实例概念条目参见表 2。

表 1. 概念条目层次扩展描述

扩展层次	包含内容
零级扩展	代表字串
一级扩展	涵义完全相同字串
二级扩展	直接相关
三级扩展	常识知识
四级扩展	下位、组成部分、相关场所、材料
五级扩展	上位（包括同一层次）
六级扩展	虚拟节点（递归扩展）

在实际应用时，以零级或一级扩展字串访问概念库，可以获得整个概念条目。为了快速地访问概念库，本文采用了哈希技术。把零级和一级扩展字串按字典序排列，并且，每个字串都可映射到相应的概念条目。概念库组织方式如图 2 所示，其中，$iL\ (0 \le i \le 6)$ 代表第 i 级扩展字串的集合。算法描述如下：

算法 1. Expand

输入：待扩展字串 input

输出：扩展后字串集合及其扩展级别

```
Begin
    用 input 查询 0L 和 1L 组成的索引，获得对应概念条目的存储地址 Address；
    在地址 Address 处读取概念条目（包括字串集合 strings 及其扩展级别）；
    if (需要六级扩展) then
        Expand(strings);
End
```

表 2. 概念条目层次扩展实例

扩展层次	包含字串
零级扩展	香港
一级扩展	香港特别行政区、香江、香海
二级扩展	港督、港币、香港经济、港澳台
三级扩展	一国两制、董建华
四级扩展	香港岛、九龙和新界
五级扩展	中国

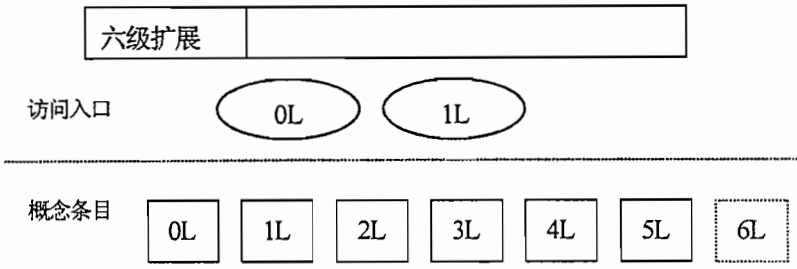

图 2. 概念库组织结构示意图

4 加权算法及句子选取

本文在实现偏重摘要时，利用了上海交通大学中英文自动摘要系统很多成果。原系统可以智能地分析 HTML、XML 和 MS WORD 等各种文件格式，获取文章的详细结构（题目、作者、各级小标题、文章段落、句子和串频等信息）。在这个基础上，进行句子加权、句子选择并最终形成偏重摘要。

4.1 句子加权

句子的权值由两部分组成：一般权值和概念扩展权值。句子的一般权值反映了句子在待处理文档和相关文档集合中的重要性。同时，句子一般权值可以在一定程度上避免偏重摘要变成一个简单字串匹配系统。句子的概念扩展权值反映了句子在多大程度上满足用户偏重要求。因此，句子的权值由一般权值和概念扩展权值共同决定。用符号 $S = T_1, T_2, \Lambda, Tc$ 表示由 C 个字串组成句子，$W_g(S)$ 表示句子的一般权值，$W_e(S)$ 表示句子的概念扩展权值。则，句子的权值 $W(S) = \frac{p \cdot W_g(S) + (1-p) \cdot W_e(S)}{\sqrt{C}}$，其中，$0 \le p \le 1$。

4.1.1 一般权值 $W_g(S)$

在自动摘要的研究开发过程中，各位专家学者提出并探讨了各种各样的方法。本文主要使用了以下几种方法：

- 标题方法

众所周知，文章的标题能够很好的揭示文章的主题。我们的统计表明：大部分科技文献（99.8%）的标题都能基本反映主题，而新闻则相对低一些（96.7%）。因此，所有出现在标题中的字串都被赋予一个值 $WofT_t$。

- 位置方法

Brandow[2]建议在摘要时提高文首段句子的权值可以提高摘要的质量。经过对大量文章的人工分析，我们发现适当提高文尾段句子的权重和语义段首句子的权重能进一步改进摘要的质量。因此，需要给每个句子乘相关的系数 $WofS_l$：

$$WofS_i = \begin{cases} a_1 & \text{首段句子} \\ 1 & \text{普通句子} \\ a_2 & \text{语义段首句子} \\ a_3 & \text{尾段句子} \end{cases} \quad \text{（其中，} a_1, a_2, a_3 \text{是常数）}$$

- 提示字串法

文章中常常有一些特殊的线索词（短语、字串、字串链），它们对文章主题具有明显的提示作用，可以利用它们来获取文章的主题。如 Edmundson[3]的文摘系统中有一个预先编制的线索词词典，词典中的线索词分为 3 种：取正值的褒义词（Bonus Words），取负值的贬义词（Stigma Words）和无效词（Null Words），文章中句子的权重为各个线索词的权重的函数。Paice[4]提出根据各种指示性短语（例如 in this paper.... the purpose of the article.....等）来选择文摘句的方法。因此，需要给每个句子乘相关的系数 $WofS_i$。

- 词频方法

能够指示文章主题的所谓有效词（或称实词）往往是中频词。根据句子中实词的个数来计算句子的权值，这是 Luhn[1]首先提出的。Oswald[5]主张句子的权值应按其所含代表性的“词串”的数量来计算；1995 年 Kenneth[6]采用相对词频的方法实现 ANES（Automatic News Extraction System）系统。假设从一篇文章中抽取 h(h 和文章的长度有关)个高频词，则它们相应的系数为 $WofT_{tf}(tf_i)$，其中，tf_i 为第 i（$1 \le i \le h$）个高频词在文中出现的频率。

于是，$W_g(S)$ 可以非常简单地根据上述因素计算出。

4.1.2 概念扩展权值 $W_e(S)$

计算概念扩展权值以前，要用概念库对用户输入的偏重字串进行概念扩展。设用户输入的偏重字串为 K，则经概念扩展后得到字串集合 $Q(q_0, q_1, q_2, q_3, q_4, q_5, q_6) = \text{Expand(K)}$。其中，$q_i(q_{i1}, q_{i2}, \ldots, q_{im})$表示第 i$(0 \le i \le 6)$ 级扩展的字串集合，该集合由 m 个字串 $q_{ij}(0 \le j \le m)$组成。设各级扩展对应的加权系数为 $Wi(0 \le i \le 6)$。

如果，句子 S 中含有第 i 级扩展字串 $C_i(0 \le i \le 6)$ 个，则该句的概念扩展权值为：

$$W_e(S) = \sum_{i=0}^{6} C_i \cdot W_i$$

4.2 摘要生成

经过上述计算，可以获得每个句子的最终权值 $W(S)$。在用户要求或理想长度下，取权值最高的句子和和相应的连贯性句子组成偏重摘要结果。本文默认摘要长度为原文有效长度的 30%或者 200 字。这个长度和 Brandow[2]建议的摘要长度基本一致。

5 实验结果

我们随机抽取了100篇汉语文章（包括科技文献、网上新闻和政府工作报告）做了实验。在用户提出偏重要求下，系统对绝大部分文章都能做出既满足用户偏重要求，又能在一定程度上反映全文主题的偏重摘要。

在摘要的评价中，由于个人主观因素占有很大比重，很难对系统做出非常客观的评价。由于教育背景、社会阅历、个人爱好等的不同，不同人对同一文章主题的理解很难一致。因此如何建立一套客观公正的评价体系和评测方法，也是一个国内外正在研究的热门课题。

6 结论和展望

本文讨论了概念库的结构，并利用概念库实现了偏重摘要系统。与偏重摘要系统的实现相似，这个概念库可以用于搜索引擎，以提高搜索引擎的查全率。除了作为一个单独系统以外，本文所介绍的偏重摘要系统也可用于搜索引擎，以方便用户准确地判断查询结果是否符合自己的需要。用于搜索引擎偏重摘要不要求很高的功能。例如，对结果不要求连贯性、不要求信息浓缩，而尽量以句子的原始形态形成结果。

从直观上分析，概念库中四~六级扩展字串用于偏重摘要系统尚可，但用于搜索引擎的查询扩展肯定会造成查询结果的泛滥，也就是查全率过高而查准率太低。由于四~六级扩展的加权系数较低，在算法实现时可以考虑用字串的同现率来代替四~六级扩展。

参考文献:

[1] H. P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 1958, 2(2): 159~165

[2] R. Brandow, K. Mitze, and L. F. Rau. Automatic condensation of electronic publications by sentence selection. Information Processing & Management 31(5): 675-685, September 1995.

[3] Edmundson H.P., New Methods in Automatic Extracting, Journal of the Association for Computing Machinery, 1969, 16(2):264~285

[4] Paice C.D. Constructing Literature Abstracts by Computer: Techniques and Prospects, Information Processing & Processing, 1990, 26(1):171~186

[5] Oswald V.A., Wyllys R.E., Automatic Indexing and Abstracting of Contents of Documents, Planning Research Corporation, Los Angeles, California, 1959

[6] Kenneth W. Church, Lisa F. Rau, Commercial Applications of Natural Language Processing, Communications of the ACM, 1995, 38(11):71~79

[7] Anastasios Tombros, Mark Sanderson, The advantages of query-biased summaries in Information Retrieval In the Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998: 2-10.

作者简介: 刘功申（1974－）,男,山东聊城人,博士生,主要研究方向为中文信息处理;胡佩华（1978－），女，河南郑州人，硕士生，主要研究方向为中文信息处理；岳奕（1978－），女，上海人，硕士生，主要研究方向为信息科学；王永成(1939－),男,江苏扬州人,教授,博士生导师,主要研究领域为中文信息处理。

A Biased Summaries and Its Implementation Technology*

LIU Gong-Shen[1], HU Pei-Hua[1], YUE Yi[2], WANG Yong-Cheng[1]

[1] *(Department of Computer Science, Shanghai Jiaotong University, Shanghai, 200030)*

[2] *(Department of Information Science, East China Normal University, Shanghai, 200062)*

E-mail:lgshen@sjtu.edu.cn

Abstract: A biased summarizes system is a significant research task. The implementation method of biased summarizes system is proposed in this paper based on a conception base which is constructed to expand conception. During the process of generating summarizes, weighting method correlated to subject and bias are discussed, and the strategy of selecting sentences to generate summarizes is discussed too. It's proved by experiment that, for a large proportion of articles, a biased summarizes which can satisfy user's biased requirement and reflect the subject of whole document can be generated by this system.

Keyword: **Conception expansion; Biased summarizes; Natural language processing;**

* Supported by the National Natural Science Foundation of China under Grant No. 60082003.