软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(5):1006−1021 [doi: 10.3724/SP.J.1001.2013.04252]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于网络的动态多文档文摘系统框架*

刘美玲[1,2], 任洪娥[1], 于 洋[1], 郑德权[2], 赵铁军[2]

[1](东北林业大学 信息与计算机工程学院,黑龙江 哈尔滨 150040)

[2](教育部-微软语言语音重点实验室(哈尔滨工业大学),黑龙江 哈尔滨 150001)

通讯作者: 刘美玲, E-mail: mlliu@mtlab.hit.edu.cn

摘 要: 在自然语言处理和计算语言学相关技术支撑下,研究基于网络的动态多文档文摘系统框架,重点描述动态多文档文摘系统框架的相关内容,介绍利用矩阵子空间方法进行动态演化建模,利用相似度和质心整体优选计算方法进行信息过滤,并利用动态流形排序方法进行句子加权的动态多文档文摘生成系统.按照多文档文摘生成步骤的划分,对 3 种创新的模型方法进行融合,综合起来从不同侧重点考虑,形成互补,提高系统性能.在网络环境下,此框架保证了动态演化的多文档文摘具有较高的信息新颖性和历史信息的演化性.

关键词: 模型框架;矩阵子空间;整体优选;动态演化

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 刘美玲,任洪娥,于洋,郑德权,赵铁军.基于网络的动态多文档文摘系统框架.软件学报,2013,24(5):1006−1021. http://www.jos.org.cn/1000-9825/4252.htm

英文引用格式: Liu ML, Ren HE, Yu Y, Zheng DQ, Zhao TJ. Web-Based dynamic multi-document summarization system framework. Ruan Jian Xue Bao/Journal of Software, 2013,24(5):1006−1021 (in Chinese). http://www.jos.org.cn/1000-9825/4252.htm

Web-Based Dynamic Multi-Document Summarization System Framework

LIU Mei-Ling[1,2], REN Hong-E[1], YU Yang[1], ZHENG De-Quan[2], ZHAO Tie-Jun[2]

[1](College of Information and Computer Engineering, Northeast Forestry University, Harbin 150001, China)

[2](Ministry of Education-Microsoft Key Laboratory of Speech Language (Harbin Institute of Technology), Harbin 150001, China)

Corresponding author: LIU Mei-Ling, E-mail: mlliu@mtlab.hit.edu.cn

Abstract: This paper introduces an Internet-based dynamic multi-document summarization system to support natural language processing and computational linguistics-related technical. This paper focuses on the description of the relevant content of dynamic multi-document summarization system framework and introduces dynamic evolutionary modeling using the matrix sub-space method, the information filtering model that uses the similarity and centroid integer selection method, and weighted sentence sorting, using the dynamic manifold method to generate the dynamic multi-document summarization system. This paper fuses the three innovation modeling methods to complement and to improve the performance of the system in accordance with the division of generated step of multi-document summarization. In a network environment, the framework ensures the dynamic evolutionary multi-document summarization with high novel information and evolutionary historical information.

Key words: modeling framework; matrix subspace; integration selection; dynamic evolution

传统的静态文摘方法把文档集看成了一个静态的文本集合,因此面对当今网络信息呈现出的较强的动态演化性和新颖性而言效果不理想,不能满足人们对网络信息获取效率的要求.动态文摘的概念是 DUC (document understand conference)[1]于 2007 年引进的,并成为 TAC2008[2]与 TAC2009 的 3 大主要评测任务之一.

* 基金项目: 国家自然科学基金(60736014, 60773069, 61073130); 国家林业行业专项(201204715)
收稿时间: 2011-12-12; 定稿时间: 2012-04-17

© 中国科学院软件研究所 http://www.jos.org.cn

在动态文摘评测任务的推动下,此研究工作获得了一定的进展.动态多文档文摘有着广阔的应用前景,可用于新闻搜索引擎、商业竞争情报分析、趋势预测等领域,通过不断满足人们的对摘要式动态信息的需求,创造更大的科学研究价值和社会价值.

本文面向网络信息动态数据流研究动态多文档文摘系统.前人的工作对动态多文档文摘的模型算法研究较少.本文着眼于从分析历史信息与当前信息关系的角度提出动态多文档文摘的建模方法,除了需要保证文摘信息的主题相关性和内容的低冗余性之外,还针对内容的动态演化性分析已出现信息和新出现信息的关系,给出问题的形式化描述,保证多文档文摘内容的动态演化性、低冗余性、新颖性、流畅性.本文重点研究以下问题:利用矩阵子空间方法进行动态演化建模,利用相似度和质心整体优选计算方法进行信息过滤,并利用动态流形排序方法进行句子加权来建立完整的动态文摘系统.综合各模型的优点,从不同侧面改善文摘系统性能.以上提到的 3 种动态多文档文摘模型各有优缺点,尝试综合起来从不同侧重点考虑,形成互补,提高系统性能.

本研究提出了对 3 种模型进行融合的完整系统框架,目前,国内动态多文档文摘的研究尚未成形,此动态文摘系统保证了动态演化的文摘具有较高的信息新颖性和历史信息的演化性,进而提高动态文摘的性能.

1 相关工作

1.1 相关研究

在多文档文摘研究领域,由美国国家标准与技术研究所(National Institute of Standards and Technology,简称 NIST)[3]组织的文本理解会议(DUC)[4]近年成为该领域相关技术发展的主要推动力.DUC 面向全球,每年都会有不同国家的不同研究机构和高等学校的研究人员提交系统,并参加 DUC.自 2001 年起,DUC 开始对多文档文摘系统进行评测,每年举行一次评测会议.在该领域的前期发展阶段,DUC 极大地推动了该领域技术的发展.由于网络信息时代的迅猛发展,网络信息铺天盖地而来,为了适应时代的需求,DUC 于 2007 年提出了一项新的评测任务,即动态多文档文摘系统.该评测任务的提出,使多文档文摘领域的研究进入了一个新阶段,使动态多文档文摘的研究成为了该领域的新热点.随着研究技术的进步,DUC 的评测规模越来越大,DUC 于 2008 年与 TREC[5]组成了一个新的评测会议,改名文本分析会议(text analysis conference,简称 TAC),以动态多文档文摘系统为评测目标的 Update Task 成为 TAC2008 与 TAC2009 的 3 大主要评测任务之一.

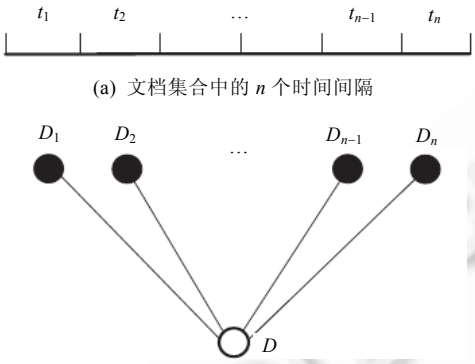
动态内容的时序划分是动态文摘的基础,相关研究在新闻事件检测(news information detection,简称 NID)[6]和 TDT(topic detection and tracking)等领域[7]受到了较多关注.时间信息在自然语言处理(natural language processing,简称 NLP)[8]领域具有非常重要的作用[9],是许多自然语言处理任务的基础,如,多文档文摘(multi-document summarization)[10]系统中需要按照时间顺序排列相关的信息;而在问答系统(question answering)中,对“何时”问句的回答更是离不开时间信息.时间信息的重要作用使得时间表达识别和规范化(temporal expression recognition and normalization,简称 TERN)[11]研究目前引起了国内外研究者们的广泛关注,国际上相关的评测有 ACE[12]和 TERN[13]等.

网络信息[14]有 3 大特点:海量性、同主题性和动态演化性.针对网络信息的 3 大特点,沈阳航空航天大学的叶娜博士[15,16]提出了基于文本主题分析的多文档文摘分析技术.该文摘技术主要是通过文本主题分析技术来解决上述多文档摘要技术中存在的问题.文本主题分析技术就是通过对文本进行浅层分析[17],识别出体现同一子主题的不同线性分割单元,对其进行主题标注并建立关联,构造文本的子主题关联图,最终实现基于语义的文本内容分析技术.该技术生成的文摘全面性广、内容冗余性低.中国科学院的许洪波博士[18,19]提出了面向 Web 话题的多文档文摘技术.该技术主要对文档集的动态演化性建模,使生成的文摘具有低的历史冗余性.

综合以上分析,基于文本主题分析的多文档文摘分析技术虽然解决了低覆盖率和高冗余性问题,但是没有捕捉到文档内容的演化性,文摘中包含大量的无用信息,这种多文档文摘方法是一种不全面的方法.面向 Web 话题的多文档文摘技术能够解决历史冗余性高的问题,使文摘捕捉到了文档内容的动态演化性,并且含有少量,甚至不包含任何历史冗余信息,文摘具有动态演化性,在某种程度上来说是一种高效的多文档文摘方法.但是该方法没有考虑到网络信息文档的同主题特征,以至于文摘不具有期望的全面性.

© 中国科学院软件研究所 http://www.jos.org.cn

动态文摘是针对网络演化信息并具有时序偏向的多文档文摘,其研究对象是网络动态演化信息的关联文档[20].如果把动态演化信息的生存时间 T 划分为 n 个时间段 $t_1,\ldots,t_n$,各个时段包含的文档集分别为 $D_1,D_2,\ldots,D_n$,则动态文摘问题可以形式化为:已知 $t_1,\ldots,t_{i-1}$ 时段的文档集 $D_1,\ldots,D_{i-1}$,求 t_i 时段的文档集 D_i 的摘要,即 $DynSummary(D_i|D_1,\ldots,D_{i-1})$,$1\leqslant i\leqslant n$,如图 1 所示.当 i=1 时,该问题退化为传统的静态文摘问题.



(a) 文档集合中的 n 个时间间隔

(b) n 个子区间对应的 D 个文摘摘要

Fig.1 Formal representation of the dynamic summarization

图 1 动态文摘的形式化表示

多文档自动文摘技术作为一种提炼概要信息的有效手段,已经得到了广泛的研究.传统的多文档文摘技术是一种静态文摘,即针对某个封闭的静态文档集生成摘要,不考虑文档集的对外联系.在 Web2.0 时代,出现在 BBS 论坛、blog、在线评论等新媒体中的网络信息(如网络话题、热点事件等,表现为一系列相关文章的集合)是动态演化的,它们随着时间的变化而出现、发展直至消亡,一个话题在不同的时刻具有不同的侧重点,而不同时刻的话题内容之间具有关联性.因此,如何从不同的侧面对动态演化的网络信息生成文摘,成为一个新的研究课题.

1.2 主流的评测方法

目前,在时序多文档文摘的评价方面完全沿用传统静态多文档文摘的评价方法,包括自动评价 ROUGE[21]、BE[22]方法和人工评价金字塔(PYRAMID)[23]方法.文摘评价主要面向文摘的内容选择和语言质量.自动评价针对文摘的内容选择进行评测,而人工评价则针对文摘的内容选择、语言质量和整体的反映度(综合考虑面向话题的覆盖度和流利度)进行评测.对于标准文摘的构建,官方有 8 个 NIST 评测者为各话题选择和撰写文摘,话题中的每个时间片均对应 4 个人工文摘.这样,人工文摘的质量将作为系统性能的上限,而基准系统的文摘(一般由文档中的首句构成)质量将作为系统性能的下限.文摘内容单元的选择和对比是文摘评价的两个关键问题.

TAC 是多文档文摘领域最有影响的国际评测会议,由美国国家技术标准局 NIS(National Institute of Standards and Technology)主办的 DUC 和 TREC 中的问答评测演化而来.TAC 评测由美国 IARPA(Intelligence Advanced Research Projects Activity)资助,每年由 NIST 的信息技术研究室中的信息检索组主办,由来自政府、企业和学术界的顾问委员会监督.Update summarization 评测面向英语,测试语料主要来自 TREC 中 QA 评测的 AQUAINT-2 数据集.

2 不同侧重点的动态模型分析

2.1 文摘系统设计灵感

动态多文档文摘的研究内容为:以历史文档集为参考,以当前文档集为研究对象,在对其动态演化性建模的基础上生成当前文档集的文摘,使其中不包含历史文档集中曾叙述的内容,最后生成的摘要随着文档集合的变

© 中国科学院软件研究所 http://www.jos.org.cn

化而动态演化,便于读者更快、更准确地获取信息,其目的是减轻对历史同主题信息已有所了解的读者获取有效信息的时间开销.为了使动态文摘系统达到这个目标,自动态文摘系统领域开创以来,各国研究者提出了多种解决方案,并且都取得了不错的效果.总结起来主要可以分为两种方法,第 1 种方法为基于文摘系统模型的改进,第 2 种为基于文摘系统生成算法的改进.

基于文摘系统模型的改进方法,主要是对传统的静态文摘系统建模方式的改进.传统的静态文摘之所以没有动态性的原因是,其研究对象是固定的文档集合,模型中没有对其信息的动态演化性进行建模.为使其具有动态性,一些研究者在传统的文摘模型中加入一个处理步骤,即历史信息过滤,形成了基于文摘系统模型改进的动态多文档文摘生成模式.历史信息过滤的主要内容为是:首先,计算当前信息与历史信息的相关度;然后,从当前文档集中过滤历史信息,得到的多文档文摘模型即为动态多文档文摘模型.

基于动态文摘系统算法的改进方法,不是对传统静态文摘模型的文摘生成阶段进行改进,而是对其特征抽取阶段和句子加权阶段的算法进行改进,以实现动态多文档文摘模型的动态性.算法改进的内容是:首先,在特征抽取阶段加入有关时间特征和历史信息特征的抽取;然后,在句子加权阶段融入对时间特征和历史信息特征的考虑,使句子的权值不仅能够反映出其内容的重要性,即内容显著度,而且能够反映出其内容新颖性,即内容新颖度.这样,生成的文摘既具有重要性又具有新颖性,使基于文摘系统文摘生成算法改进的文摘系统具有了动态性.

以上两种动态演化性建模方法各有优缺点:基于模型的改进方法的优点在于把当前文档集中的历史信息过滤掉了,对历史信息的处理比较彻底,所以具有非常好的动态性,但是其在过滤历史信息的同时也过滤掉了不少当前的重要信息,所以也减弱了文摘的显著性;基于算法的改进方法其动态演化性建模是基于特征抽取和句子加权的,所以其存在对历史信息处理的不彻底的缺点,但是保证了文摘的显著性优势.为了使动态多文档文摘系统在保证文摘显著性的同时又具有很好的动态性,本文提出了一种将上述两种动态演化性方法进行结合的方法,即基于文摘系统模型和算法同时改进的动态文摘模型.

2.2 各个侧重模型的特色

子空间模型、基于相似度和质心整体优选的文摘句过滤模型和动态流形排序模型是本文提出的 3 个动态文摘模型,都用于动态多文档文摘领域中,主要用来对相关文档集的动态性演化性进行建模,使文摘具有动态性.各模型中使用的思想和算法不同,因而所产生文摘的效果也不同.3 个模型的侧重点各异:

- 子空间模型从整体入手,侧重的是相应文档集的整体信息空间,把历史文档集和当前文档集分别信息化为历史信息空间和当前信息空间.子空间模型方法运用矩阵空间理论方法,找出历史信息空间和当前信息空间的相交空间和相异空间,然后通过过滤相交空间保留相差空间的方法,达到对动态演化性建模的目的.
- 基于相似度和质心整体优选的句子过滤模型主要以句子为研究对象.利用相似度来衡量句子之间的信息重叠度,通过过滤掉与历史文摘中句子信息重叠度大的句子来达到冗余信息过滤的目的,使文摘中不含历史冗余信息,具有新颖性.

上述两种模型的一个最大缺点是,把文档集中的所有句子都孤立化了,认为句子之间没有任何联系,这种假设过于理想,将影响文摘的生成效果.

- 动态流形排序模型弥补了上述两种模型的缺点.此模型的研究对象是句子之间的关系.根据相关句子之间的信息传递性原理,利用句子间的相关性对句子进行排序.排序后,相似的句子趋向于具有相同的排序值,所以文摘中不会漏掉任何重要性高的句子,也不会误选任何低重要性的句子.而且融入对句子历史信息特征的惩罚和时间特征的奖励后,还能实现对文档集所含信息动态演化性的建模,使文摘具有动态性.

综上所述,3 种动态多文档文摘模型各有优缺点,如果能够互补,将使文摘效果更好.所以,本文进一步的研究工作就是研究如何对 3 种模型进行融合,使文摘具有更好的显著性和新颖性.

© 中国科学院软件研究所 http://www.jos.org.cn

3 动态多文档文摘系统

3.1 系统模块设计和划分

本文使用基于文摘系统模型及算法同时改进的动态文摘模型对文档集信息的动态演化性进行建模,因此,融合 3 种创新模型的特点,此动态多文档文摘系统首先应该包括信息过滤模块,遵循传统多文档文摘系统处理核心,所以必须包含传统静态多文档文摘系统的相关子模块,即文档集预处理模块、特征抽取模块、句子加权模块、文摘句选择及排序模块.最后,还必须对其特征抽取模块及句子加权模块进行算法改进,使其具有动态性.

综合起来,本文提出的动态多文档文摘模型主要由文档集预处理模块、特征抽取模块、信息过滤模块、句子加权模块及句子选择及排序模块组成.系统流程图如图 2 所示,其中,深色虚线框区域为主模块.

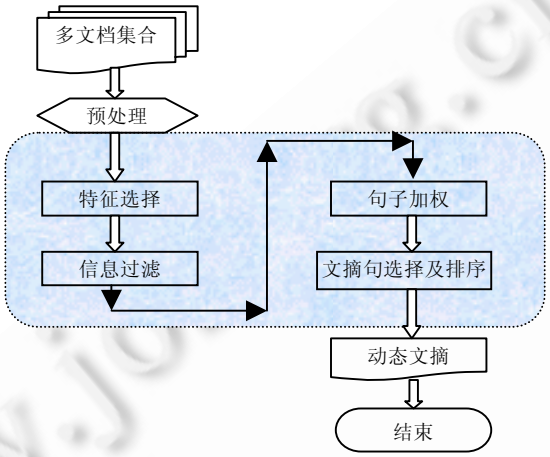


Fig.2 System module flow chart

图 2 系统模块流程图

由于本系统的主要创新性研究内容为对文档集的动态演化性建模,所以本文重点介绍特征抽取模块、信息过滤模块及句子加权模块.其中,预处理模块使用本实验室开发的预处理系统,文摘句选择及排序模块使用 MMR 方法[24]和 MO 方法[25].

3.1.1 特征抽取模块

为使此动态文摘系统除了具有传统静态多文档文摘系统所具有的性能之外还有独特的动态性,本动态文摘模型的特征抽取模块除了包含传统文摘系统特征抽取模块中的主题词抽取、长度特征抽取、位置特征抽取等模块以外,还必须引入句子的显著性特征抽取、历史信息特征抽取及时间特征抽取模块.句子的显著性特征是为了进一步提高本动态文摘系统相应于传统静态多文档文摘系统各项性能而引入的,而历史信息和时间特征是为了使其具有且提高其动态性而引入的,其结构如图 3 所示.

(1) 主题词抽取及其权值计算

主题词抽取的过程为:首先统计文档集中所有词语在该文档集中的出现频率,即词频 TF;然后,从词语集合中删除所有停用词,对词语集合中剩余的词语从有序的词语集合中抽取出指定数目的词语组成另一个词语集合,即为主题词集合.这是常用的 TF-IDF[26]主题词提取算法,由于系统后面各个模块的需要,本系统除了把停用词之外的所有词语都作为关键词以外,在主题词特征抽取子模块,本文还提出了一种新方法,即 TF-IDF-ISF 主题词权值计算方法.在系统实现阶段将详细加以介绍.

(2) 句子历史冗余性特征值的计算

句子历史冗余性特征即为句子所含历史信息的度量,此特征是动态多文档文摘系统区别于传统文摘系统的重要度量.动态多文档文摘系统之所以具有动态性特征,是因为动态文摘方法能够通过比较当前信息和历史

© 中国科学院软件研究所 http://www.jos.org.cn

信息的相同点和不同点来刻画信息的动态演化性,然后根据句子的动态演化性对句子加权,使生成的文摘具有动态性.为了实现此目的,本系统通过句子的历史冗余性来刻画句子所含信息的动态演化性.本系统所用的句子历史冗余性特征值的计算方法如下:首先对历史文档集进行处理,生成历史文摘;然后,以当前文档集的句子集合为研究对象,且以历史文摘为参考,计算当前句子集合中每个句子所含内容与历史文摘所含内容的信息重合度,此信息重合度即为句子的历史冗余性特征值.

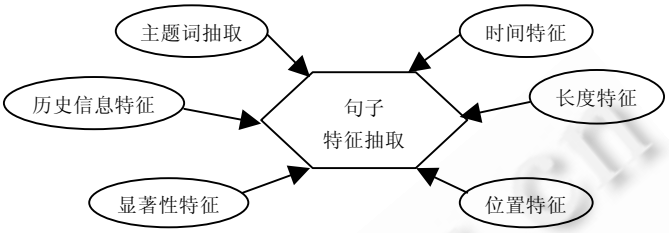


Fig.3 Feature extraction structure diagram

图 3 特征抽取结构图

(3) 句子显著性特征值的计算

句子显著性特征即该句子所含信息对其文档集的代表性,因此,句子的显著性特征值越大,则意味着其成为文摘句的可能性越大,并且此句子的加入将提高文摘的相应评测打分,最终提高该动态多文档文摘的性能.根据数学上的投票原理,即某一事物得到其他事物的认同率越高,说明该事务相对于其他事物的重要性权重也就越大.根据此原理,文档集句子集合中某一句子比集合中其他句子的认同率越高,则说明该句子越重要,即其显著性越大.这种句子对句子之间的认同率就是句子之间的相似度.综上所述,得出一个重要的结论:句子与文档集句子集合中所有其他句子的相似度累加值是句子重要性的一个度量,称其为句子显著度.

(4) 句子时间特征值的计算

文档句子的时间特征抽取是动态多文档模型的一个难题,这是时序多文档文摘的主要研究内容,涉及到时间表达式的识别抽取与归一化,研究的工作量以及复杂性都与本文不同,不属于本系统的研究范畴.因此,本系统根据国际评测语料的时间特性,研究对文档集中各个句子的时间特征进行宏观抽取.其具体方法是:首先,确定文档集中所有文档的发表时间;然后,根据文档的发表时间按时间先后顺序对文档集中所有文档进行排序,确定其时间特征排序值,并统计文档集中文档的总数量;最后,根据文档在文档集中所有文档中的排序值确定其中所有句子的时间特征权重.本文拟使用的具体方法是:把文档的时间特征排序值的倒数值确定为其所含所有句子的时间特征排序值.

(5) 句子长度特征值的计算

为了提高该系统在实际工程中的应用性,文摘系统所生成的文摘应符合阅读和理解的要求.为此,文摘中所有句子都应满足一定的长度限制.文摘系统的目的是使文摘中包含尽可能多的重要信息,使读者花费较少的时间代价获取尽可能多的信息.这样,文摘系统就要求其文摘中的每个句子都不能太长.因为长句子即使能够包含很多重要信息,但是由于其占用过大的文摘空间,会导致信息空间比低,这样就使得文摘中总信息量下降.类似地,太短的句子通常会包含较少的信息量,同样会导致整个文摘中包含的信息量较少,所以文摘中也应尽量不包含太短的句子.基于以上分析,本系统在对句子的长度特征进行抽取时应对抽取的所有句子设置一定的长度限制,使文摘中含有尽可能多的信息量.

(6) 句子位置特征权值的计算

据语言学统计研究表明,文档中句子的分布具有一定的特征,大部分重要度较大的句子一般来说都位于文档的开头以及结尾,文档中间部分大量的句子都属于解释性及叙述性的句子,其信息量相对于开头部分及结尾部分的句子较少.既然动态文摘系统的目的是抽取高信息量的文摘句组成候选句子集合,因此在传统多文档系

© 中国科学院软件研究所 http://www.jos.org.cn

统中特征抽取模块中将其视为文摘句的一个重要特征,在该动态多文档文摘系统中同样将其视为文摘句特征抽取的一个重要特征.

3.1.2 信息过滤模块

动态文摘系统的处理对象为当前文档集.此文档集中包含了一部分历史冗余内容,这部分内容在历史文档集中已有所陈述,对读者而言它属于垃圾信息,所以就没有必要花费时间和精力去了解它.因此,设计信息过滤模块进行对历史文档集和当前文档集之间的关系分析如图 4 所示.

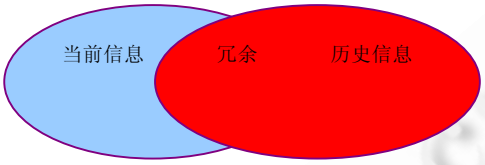


Fig.4 Current document and historical documents diagram

图 4 当前文档与历史文档关系图

其中,在当前文档集中包含一部分历史冗余信息,为的是使文摘体现出动态性,在文摘系统中必须删除这部分冗余信息,这就是动态文摘系统中信息过滤模块需要解决的问题.

信息过滤模块是该动态文摘系统实现动态性的关键步骤之一,因为它对原始句子集合进行整理,对当前文档集中的历史信息进行过滤,滤掉了原始句子集合中包含历史信息大的句子,使集合中所剩句子都为具有动态性信息的句子.

这样处理过的句子集合就成为动态句子集合,使其成为本系统后续模块的研究对象,这就保证了本系统结果文摘的动态性.动态性得到了保证,但是又出现了另外一个问题,即该模块对当前文档集句子集合的处理是否会减弱结果文摘的全面性和代表性,答案是肯定的.为了解决此问题,本模块在对句子集合处理时,不仅要保证结果文摘具有动态性,而且要保留其原有的全面性以及代表性.所以,本信息过滤模块在对信息进行过滤时限制了过滤掉的信息的数量,其实现方式为限制从原始句子集合中过滤掉的句子的数量.

3.1.3 句子加权模块

经信息过滤模块处理后,当前文档集如图 5 所示.深色部分即为处理后的句子集合,其中不再含有历史冗余信息,因而体现了文档集合的动态变化,成为本模块的处理对象.

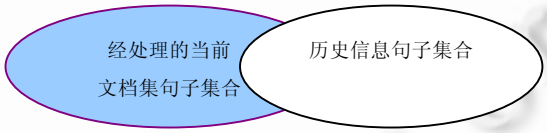


Fig.5 Information filtering

图 5 信息过滤

在设计本模块之前,先简单分析一下其他动态文摘系统的主要特点.总体上来说,目前的动态文摘系统为体现动态性,主要有两种解决方案,分别在两个不同的处理模块中实现:一是在常规的文摘系统的加权模块之前插入一个信息模块,目的是在此模块中通过信息过滤的方法过滤掉当前文档集句子集合中包含历史信息大的句子,把原始句子集合转换为动态句子集合,通过后续相应阶段的处理即可生成具有动态演化性的文摘;二是通过对常规文摘方法的句子加权模块加以改进,以达到使结果文摘具有动态演化性的目的.例如,有的动态文摘方法在句子加权阶段加入一项能够反映动态性的句子特征,比如说对具有历史冗余性的主题词给予惩罚值、对具有信息新颖性的主题词给予奖励值,都能把句子的动态性特征引入到其权值当中,这样也能使结果文摘具有动态性.

这两种动态性的实现方法各有优缺点.第 1 种方法在删除句子数量足够多的前提下,能够使结果文摘具有

© 中国科学院软件研究所 http://www.jos.org.cn

很强的动态性,但是它也可能会破坏文摘的全面性和代表性,这是我们所不希望看到的.若删除的句子数量太少,文摘将不具有理想的动态性.第 2 种方法的主要优点是在保证文摘动态性的前提下,还保留了文摘原有的全面性和代表性,其缺点是结果文摘动态性的强度不理想.

通过以上分析可以看出,上述两种方法是互补的,其结合能弥补相互之间的缺点,而且其优点为其二者优点的结合.受此启发,本系统设计了一种结合上述两种方法的动态文摘方法进行句子加权.首先,通过信息过滤的方法对原始句子集合进行处理,使其变成动态句子集合.为了不破坏结果文摘的全面性和代表性,本系统把应过滤的句子数量限制在一定的范围之内,这样就既保证了文摘的动态性,又不至于破坏其全面性和代表性.文摘阅读者可能会怀疑这样生成的文摘的动态性强度不理想,这也是本文前面部分所提过的第 1 种方法的缺点.本系统通过在加权模块中结合第 2 种方法来解决此问题,提高文摘的动态性.在加权阶段如何结合上述第 2 种方法,正是本模块要讨论的主题.

流形排序[27]思想是一种理想的句子加权算法.它主要是在如下两项假设的基础上发展起来的:

1) 邻近的结点很可能有相同的排序分数;

2) 在相同结构(一个类或簇)上的结点很可能有相同的排序分数.

可以看出,假设 1)属于局部范围的一致性假设,而假设 2)属于全局范围的一致性假设.流形结构中结点的排序取决于局部和全局上下文信息,具体到句子加权模型中,其意思是:在句子集合中,内容相似度大的句子其权值也应该相同.因此,利用流形排序算法对句子加权,将能够为集合中的每个句子赋予确切的权值,此权值能够正确地反映出该句子对文档集信息的重要性以及其成为文摘句的可能性值.传统的流形排序算法仅能应用于静态文摘系统中,为静态文档集句子集合中的句子赋予权值,不具备融入句子动态特征的能力.为了能够在动态文摘系统中利用流形排序算法为句子赋予权值,本模块设计了一种动态流形排序算法,旨在为句子集合中的句子赋予融入动态特征的权值.

动态流形排序算法主要包括 3 个步骤,如图 6 所示,即句子相似度矩阵建立、句子初值计算以及排序值的迭代计算.

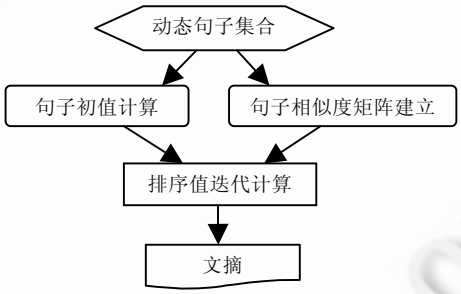


Fig.6 Dynamic manifold sort

图 6 动态流形排序

3.2 动态文摘系统框架的实现

本文在之前的基础实验中提出了 3 种新的句子加权方法[28]:短语的信息粒度表示方法、动态 TF-IDF-ISF 的句子加权方法和句子相似度计算的句子加权方法.对基于矩阵子空间分析的动态文摘模型而言,是运用相应的静态文摘方法对处理后的句子集合进行处理以生成文摘.本节实验了上述 3 种加权方法,并对实验结果进行了测试,验证了该动态模型和相应句子加权方法的有效性.同时,在第 2 种动态模型方法上也实验了这 3 种句子加权方法,依然获得了良好的效果.

动态多文档文摘模型的重点在于实现系统动态性.其他系统性能虽说也重要,但是其有传统静态多文档系统的实现方法作为基础,实现起来比较容易,因此不是该模型的重点,所以本节只对其作简要的介绍.本节将对该模型实现动态性的模块和算法的关键问题及其解决方案做重点介绍.

© 中国科学院软件研究所 http://www.jos.org.cn

下面介绍各个模块具体的设计实现方法,并给出关键问题的解决方案.

3.2.1 特征抽取模块的实现

(1) 主题词抽取方法

传统的主题词提取方法是基于 TF*IDF 的主题词提取方法.它考虑了词语本身信息即词频及词语所属文档集信息对主题词的影响,但是没有考虑到词语所属句子集合信息对其的影响.所以,为了进一步提高该动态文摘模型的传统文摘性能,本节提出了一种新的主题词抽取方法,即基于 TF*IDF*ISF 的主题词提取方法.在其中加入该词语所属句子集合的信息对其的影响,其中,TF 为相应词语在文档集中的词频;IDF 为相应词语的反文档频率,其计算方法为:首先确定在文档集中包含该词语的文档的数目,然后计算其倒数,该倒数即为该词在文档集中的反文档频率 IDF;ISF 为词语在文档集中的反句子频率,其计算方法为:首先确定在句子集合中包含该词语的句子的数目,然后计算其倒数,该倒数即为该句子的反句子频率.相应词语的词频、反文档频率和反句子频率确定之后,可通过公式(1)计算词语 w 的权值.

$$Wgt(w)=TF(w)\times IDF(w)\times ISF(w) \tag{1}$$

待文档集中所有词语的相应权值计算完毕之后,可根据其权值确定词语的重要性,继而确定文档集的主题词集合.

(2) 句子历史冗余性特征值的计算

句子历史信息特征值的计算是本节的一个重点.它是使该动态多文档文摘模型具有动态性的特征之一.本系统主要通过所研究句子与历史文档集句子集合之间的关系来确定其历史信息特征值.主要方法是:首先计算该句子与历史文档集句子集合中所有句子的相似度,然后累加;因为其累加值能够反映出其与历史信息的内容相关度,自然可以衡量该句子的历史信息特征,所以,本系统以其作为句子的历史信息特征值.其计算公式如下:

$$NWgt(s)=\frac{\sum_{i=1}^{m}\left(\frac{\sum_{j=1}^{n}Wgt(w_j)}{length(s_i)}\right)}{length(s)\times count} \tag{2}$$

其中,$NWgt(s)$即为句子 s 的历史冗余性特征值,m 为历史文摘中文摘句的总数,n 为句子历史文摘中句子 s_i 的同现主题词数量,$Wgt(w_j)$为主题词 w_j 的权重(详见公式(1)),$length(s_i)$和 $length(s)$分别为句子 s_i 与句子 s 中的主题词词语总数,$count$ 为历史文摘句子集合中句子的总数.

(3) 句子显著性特征值计算

本系统通过计算相应句子与所有其他句子的相似度累加和来计算句子的显著性特征值,其计算公式如公式(3)所示:

$$SWgt(s)=\frac{\sum_{i=1}^{m}\left(\frac{\sum_{j=1}^{n}Wgt(w_j)}{length(s_i)}\right)}{length(s)\times count} \tag{3}$$

其中,$SWgt(s)$即为句子 s 的显著性特征值,m 为文档集中句子的总数,n 为句子 s_i 与句子 s 中同现的主题词总数,$Wgt(w_j)$为主题词 w_j 的权重(详见公式(1)),$length(s_i)$和 $length(s)$分别为句子 s_i 与句子 s 中的主题词词语总数, $count$ 为当前文档集句子集合中句子的总数.

(4) 句子时间特征值计算

句子的时间特征是继句子历史信息特征之后的又一动态文摘系统的重要特征,是为了进一步提高该动态多文档的动态性能而引入的特征,如前所述,该模型主要通过句子所属文档在文档集中的时间排序值来确定其时间特征.此思想能够以较小的时间及空间复杂度来提高系统的动态性,是一种高效的特征抽取方法.根据此思想,句子的时间特征计算公式如下:

© 中国科学院软件研究所 http://www.jos.org.cn

$$TWgt(s)=1/n \tag{4}$$

其中,n 代表按照发表时间排序后的文档集中句子所属文档的排序值.

(5) 句子长度特征值计算

由前文相关内容的陈述可知,文摘句过长和过短都不利于文摘系统性能的提高,所以本系统在特征抽取模块中加入对相应句子长度特征的抽取,即可在句子加权模块中融入基于句子的长度特征的权重,最终达到提高文摘系统性能的目的.本系统通过对长句和断句进行惩罚的方法来限制文摘中句子的长度,具体惩罚公式如下:

$$LWgt(s)=\begin{cases}1/(Length(s)-0.5\times \mathrm{Max}Length), & Length(s)>0.5\times \mathrm{Max}Length\\ 1/(0.5\times \mathrm{Max}Length-Length(s)), & Length(s)\leqslant 0.5\times \mathrm{Max}Length\end{cases} \tag{5}$$

其中,$Length(s)$表示句子 s 的长度,$LWgt(s)$表示句子 s 的长度权重,Max$Length$ 表示文档集合中句子的最大长度.

(6) 句子位置特征值计算

句子长度特征是文摘系统最常用的特征,由于其思想及实现算法已相当成熟,本系统无须再对其进行改进,将直接应用经典的实现算法.本系统通过以下公式计算句子的位置权重值:

$$PWgt(s)=1/n \tag{6}$$

其中,n 代表句子 s 在其所属文档中的位置值.

3.2.2 信息过滤模块的实现

其主要处理算法是:首先,根据句子的历史冗余性特征值对当前文档集句子集合中的所有句子按从高到低进行排序,排序靠前的句子具有较大的历史冗余性特征值,意味着其与历史信息具有高的信息重合度,其信息动态性低,应对其作删除处理.从句子集合中删除指定数量的句子,集合中所剩的句子即为信息动态性相对较大的句子.

处理结束后,集合中所剩句子重组的集合即为动态句子集合,为本系统后续模块的处理对象.需要注意的是,所应过滤的句子数量的确定问题是一个难以解决的问题,因为若过滤的句子数量太少,则其结果文摘不具有理想的动态性;若过滤的句子数量太多,将不足以保留结果文摘的全面性和代表性.因此,其数量的确定应做到恰到好处,视具体情况而定.本系统通过一系列测试设定的经验值为 50 句左右.

3.2.3 句子加权模块的实现

句子加权模块由动态流形排序算法 3 个步骤完成,下面给出具体描述.

(1) 句子相似度矩阵的建立

句子相似度矩阵的建立是动态流形排序算法的基础,其关键是句子相似度计算方法的研究.目前,此项研究工作已相当成熟,已有多种计算方法相继出现,并且均获得了良好的性能.但是,其在文摘系统中的应用效果不是很理想.因此,根据文摘系统的实际情况,结合本系统特征提取模块主题词权值的计算,本系统提出了一种新的句子相似度计算公式.实验结果表明,其在文摘系统的应用中获得了不错的效果.其主要计算如公式(7)所示:

$$Sim(s_i,s_j)=\frac{\sum_{k=1}^{count} Wgt(w_k)}{length(s_i)+length(s_j)} \tag{7}$$

其中,$Sim(s_i,s_j)$为句子 s_i 和 s_j 相似度;$count$ 为句子 s_i 和 s_j 同现主题词的数量;$Wgt(w_k)$为主题词 w_k 的权值(见公式(1));$length(s_i)$和 $length(s_j)$分别为句子 s_i 和 s_j 的长度,即所含主题词的数量.

根据公式(7)计算当前文档集句子集合中所有句子两两之间的相似度值,即可构成一个 $n\times n$ 相似度矩阵 W;n 为集合中句子的总数量,其元素 $W_{i\times j}$ 即为句子 s_i 和 s_j 的相似度值.注意,此相似度矩阵为一个对称矩阵,为了避免迭代过程中句子权值的增强,本系统设置元素 $W_{i\times i}$ 的为 0.

(2) 句子初值的计算

句子初值的计算是动态流形算法的重要步骤之一.它体现了一个句子的原始重要性.本系统所应用的赋初值公式如下:

© 中国科学院软件研究所 http://www.jos.org.cn

$$FWgt(s)=\alpha\times\sum_{i=1}^{count(s)}Wgt(w_i)+\beta\times LWgt(s) \tag{8}$$

其中,$count(s)$为句子 s 的长度;$Wgt(w_i)$为主题词 w_i 的权值;$LWgt(s)$为句子 s 的长度特征值(见公式(4));α和β分别为参数,其经验值为α=0.2,β=0.8.

在句子初值计算公式中融入主题词权值的原因是,其可以体现句子在句子集中的重要性,最终提高文摘的代表性和全面性.而融入句子的长度特征即可控制文摘句的长度,削弱长句和短句选为文摘句的可能,使文摘获得更好的性能.

(3) 排序值的迭代计算

此步骤为动态流形排序算法的关键步骤,因为在该步骤中不仅加入了句子的重要度特征值,而且加入了句子的动态特征值.传统的排序值迭代公式如下:

$$f(t+1)=\alpha\times S\times f(t)+(1-\alpha)\times y \tag{9}$$

其中,$f(t+1)$为一次迭代计算后的句子权值向量,$f(t)$为一次迭代计算前的句子权值向量,S 为相似度矩阵,α和 $1-\alpha$ 分别表示相邻结点和初始的查询数据点的排序值对当前排序值的相对贡献.

为了体现动态流形排序算法的动态性,本系统在其迭代公式中同时融入了体现动态性的特征和体现显著性的特征,提出的句子排序值迭代计算公式如下:

$$f(t+1)=\alpha\times TWgt(s)+\beta\times PWgt(s)+\eta\times SWgt(s)-\mu\times NWgt(s)+\gamma\times S\times f(t) \tag{10}$$

其中,$TWgt(s)$为句子 s 的时间特征(见公式(3));$PWgt(s)$为句子 s 的位置特征值(见公式(2));$SWgt(s)$为句子 s 的显著性特征值(见公式(5));$NWgt(s)$为句子 s 的历史冗余性特征值(见公式(6));$\alpha,\beta,\gamma,\eta,\mu$分别为参数;$S$ 为相似度矩阵;$f(t+1)$为一次迭代计算后的句子权值向量;$f(t)$为一次迭代计算前的句子权值向量.注意,还应根据实际情况确定相应的迭代次数.

由特征抽取的分析可知,句子的时间特征和历史冗余性特征主要用来体现句子的动态性.因此通过这两项特征,可削弱历史冗余性大的句子的权值,同时增强新颖性大的句子的权值.而句子的位置特征值、显著性特征值以及其初始权值能够充分体现该句子对文档集信息的代表性和概括性,并且通过动态流形排序算法可以使内容相似的句子之间进行权值的相互渗透,如此,即可消除因随机因素而使重要性和动态性大的句子的权值低的问题,同时也消除了因随机因素而使重要性和动态性小的句子的权值高的问题.即,使文摘拥有了动态性,同时还保留了文摘应有的代表性和全面性,提高了文摘的性能.

3.2.4 候选句生成模块

本系统提出了一种改进的 MMR 去冗余算法.它充分考虑到了句子之间在内容上的冗余性.

该算法是基于主题词权值建立起来的,其计算公式如下:

$$AZWgt(s)=\alpha\times BZWgt(s)-\beta\times\sum_{i=1}^{n}\frac{\sum_{j=1}^{simcount}Wgt(w_j)}{\sum_{k=1}^{count(s_i)}Wgt(w_k)} \tag{11}$$

其中,$AZWgt(s)$为改进去冗余算法处理前的候选文摘句 s 的权值;$BZWgt(s)$为改进去冗余算法处理之后的候选文摘句 s 的权值;$Wgt(w_j)$和 $Wgt(w_k)$分别为主题词 w_j 和 w_j 的权值(见公式(1));n 为文摘句集合中的句子数量;$Simcount$ 为候选文摘句 s 和文摘句 s_i 同现的主题词的数量;$count(s_i)$为文摘句 s_i 的主题词总数;α和β分别为参数,其具体值应视具体情况而定,其经验值为α=0.3,β=0.7.

4 实验结果与分析

4.1 实验数据及评价

在 DUC2007 中,Update Summarization 作为一个先导任务,测试语料从主要任务的 45 个话题中挑选 10 个话题,每个话题由 3 个连续进化的时间片组成,即每个话题有 3 个时序相关的文档集 A,B,C,分别包含 10,8,7 个文

© 中国科学院软件研究所 http://www.jos.org.cn

档.假定读者对先前的文档有一定的了解,Update Summarization 任务的目的就是针对每一个时间片给出 100 字的文摘,该文摘反映沿着时间线的内容进化.可见,文摘的发展方向在发生变化.

在 TAC2008 中,Update Summarization 任务的测试语料由来自 AQUAINT-2 的 48 个话题组成,每个话题包含两个时间片,且均由 10 个文档组成.例如,话题“D0801A”由两个时间片“D0801A-A”和“D0801A-B”组成,二者内部的 10 个文档分别由各自的 id 来表示,话题本身由⟨title⟩和⟨narrative⟩来描述.

评价标准采用文摘评测领域著名的 ROUGE 工具,其中最主要的两个指标 ROUGE-2 和 ROUGE-SU4*是评价 Update 文摘的.本文在 TAC2008 的 Update Summarization 测试数据上,将动态文摘结果的 ROUGE-2(R-2)和 ROUGE-SU4*(R-SU4*)得分与 TAC 2008 Update 实际系统的得分进行了对比,结果表明,本文的动态多文档文摘方法具有良好的性能.

4.2 实验结果

(1) 参数测试

动态流形排序算法中,各特征项对文摘性能的影响很大,因此,本文第 1 组实验由若干小实验组成.本算法的参数比较多,参数调整的空间比较大,通过调整可以得出更好的结论,有助于提高系统的潜在性能.本系统在 ROUGE 评测系统体系下对算法参数的调整过程如图 7~图 11 所示.

图 7 显示了该动态文摘模型中时间特征值对系统性能的贡献.观察可知,当其贡献参数值大于等于 0.35 时,系统的性能最好,即该特征值对系统性能的影响最大.尤其是 ROUGE-SU4,随着该参数值的变化其变动很大,表明该特征与系统动态性能之间的联系很紧密.

图 8 展示了动态多文档文摘模型中句子的位置特征值对其性能的影响.观察可知,当其贡献值参数β在 0.15~0.2 之间时,系统性能最好,表明句子的位置特征是该多文档文摘模型不可忽视的一部分.

图 9 显示了流形排序算法迭代计算前的句子权值对系统性能的影响.当其贡献参数γ大于等于 0.35 时,系统的性能最好,表明其对系统的总体性能的贡献比较大,是系统设计时一个非常重要的部分.

句子的显著性特征是本动态多文档文摘模型为提高系统的传统文摘性能而引入的一个特征,其对系统性能的影响如图 10 所示.当其贡献参数值为 0.15 左右时性能最好,表明此特征的加入在不占用系统太多贡献值的情况下能够在一定程度上提高系统的性能,是一种性价比较高的特征.

句子的历史信息特征是为提高系统的动态性能而加入的一项特征,其对系统动态性能的影响可通过对图 11 的观察得知.当参数的值为 0.1 左右时,系统的性能达到最好.从图 11 还可以看出,随着参数值的变化,两个性能指标的值变化比较统一,表明参数对系统的总体性有所影响,在其值设置恰当的基础上,能够较大程度地提高系统的动态性能.

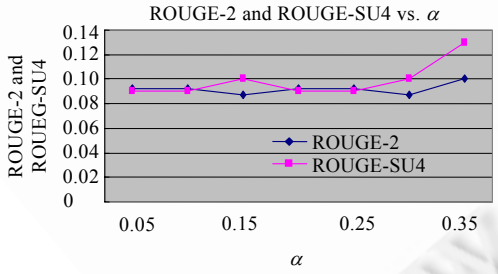


Fig.7 Dynamic model parameter α impact on system performance

图 7 动态文摘模型中参数α对系统性能的贡献

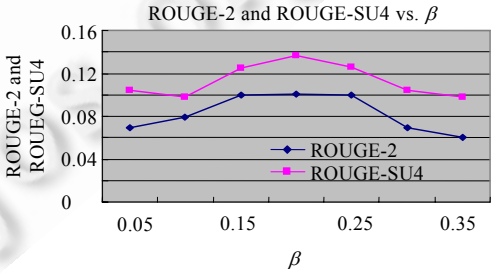


Fig.8 Dynamic model parameter β impact on system performance

图 8 动态文摘模型中参数β对系统性能的贡献

© 中国科学院软件研究所 http://www.jos.org.cn

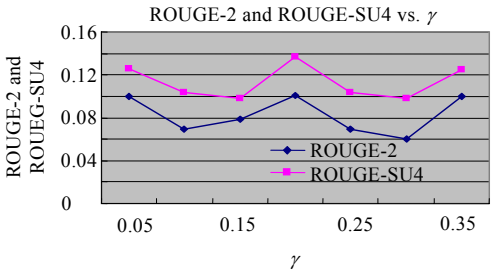


Fig.9 Dynamic model parameter γ impact on system performance

图 9 动态文摘模型中参数γ对系统性能的贡献

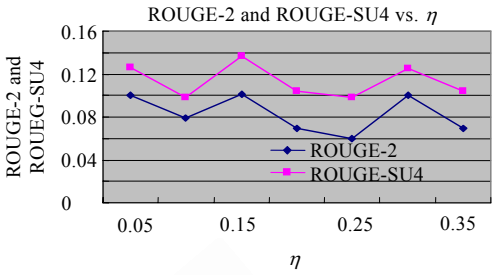


Fig.10 Dynamic model parameter η impact on system performance

图 10 动态文摘模型中参数η对系统性能的贡献

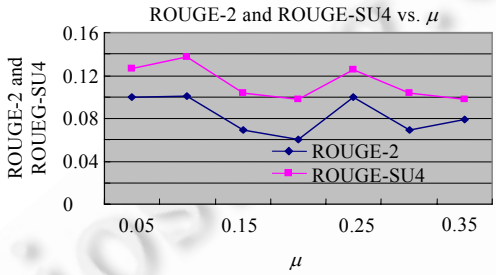


Fig.11 Dynamic model parameter μ impact on system performance

图 11 动态文摘模型中参数μ对系统性能的贡献

最终,句子加权结果比较见表 1,Experiment 5 显示最好的系统性能.

Table 1 Features effect on summarization performance in the manifold sorting algorithm

表 1 流形排序算法中各特征项对文摘性能的影响

Experiment	α	β	γ	η	μ	R-2	R-SU4*
Experiment 1	0.2	0.2	0.2	0.2	0.2	0.069	0.104
Experiment 2	0.1	0.1	0.2	0.3	0.3	0.079	0.098
Experiment 3	0.3	0.2	0.3	0.1	0.1	0.100	0.125
Experiment 4	0.3	0.1	0.3	0.1	0.2	0.069	0.104
Experiment 5	0.2	0.1	0.2	0.3	0.2	0.101	0.137
Experiment 6	0.1	0.1	0.2	0.4	0.2	0.060	0.098
Experiment 7	0.1	0.2	0.3	0.3	0.1	0.100	0.126

(2) 信息过滤算法中参数对文摘性能的影响

第 2 组实验是对系统参数 n,即信息过滤模块中得到的最佳候选句子数进行调试,见表 2,主要是为了找出信息过滤模块中应过滤的句子的最佳数量.从 5 个有代表性的实验结果中可以看出,当 n=100 时,即 Experiment 3 从文档集中过滤掉 100 句包含历史信息大的句子时,文摘系统的性能达到最好,过滤掉过多的信息或者过少的信息都会影响系统的性能.

Table 2 Number of filter sentences effect on system performance in the information filtering model

表 2 信息过滤模块中过滤句子数对系统性能的影响

Experiment	n	R-2	R-SU4*
Experiment 1	10	0.074	0.098
Experiment 2	50	0.100	0.125
Experiment 3	100	0.101	0.137
Experiment 4	130	0.065	0.101
Experiment 5	150	0.065	0.101

© 中国科学院软件研究所 http://www.jos.org.cn

上述分组实验验证了如下结论:当信息过滤模块中过滤的信息过少时,将不足以过滤掉当前文档集中的历史信息,这样就影响到了系统的动态性能;反之,当过滤掉的信息过多时,将不可避免地过滤掉当前文档集的一部分有用的信息,也就减弱了文摘的概括性,同样会影响文摘的总体性能.因此,信息过滤模块是本系统的一个不可或缺的部分.

(3) 结果比较

TAC2008 不仅为广大研究者提供了动态文摘系统的实验语料,而且提供了国际上最新的动态文摘系统的评测得分作为参考,供研究者们对比,方便其研究工作,以促进动态文摘技术的发展.同样,本系统以 TAC2008 提供的高性能系统的评测得分作为参考,定位本动态文摘的性能,以便于进一步改善本系统.本系统与国际评测系统的打分比较如图 12 所示.

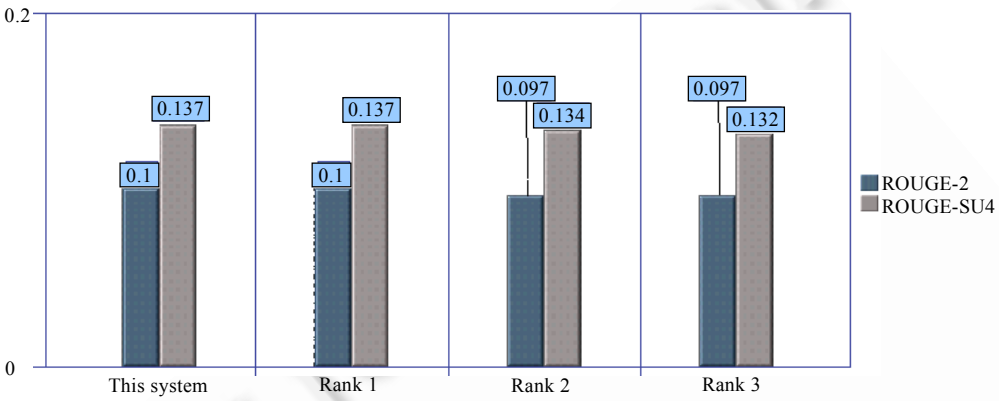


Fig.12 Dynamic system compares to the TAC2008 actual systems

图 12 动态系统与 TAC200 系统打分的比较

从图 12 中可以看出,本系统的得分已经超过了 TAC2008 中第 2 名与第 3 名系统的评测得分,与第 1 名系统得分相同,说明本动态文摘模型在国际的动态文摘领域已经处于前沿位置,是一种高性能的动态文摘模型.本文还将本模型做后续研究,以更好的成绩参加国际标准评测.

5 结 论

本文介绍了利用矩阵子空间方法进行动态演化建模,利用相似度和质心整体优选计算方法进行信息过滤,并利用动态流形排序方法进行句子加权的文摘生成系统.按照多文档文摘生成步骤的划分,从不同侧面改善文摘系统性能.3 种动态多文档文摘模型结合,形成互补,保证了动态演化的文摘具有较高的信息新颖性和历史信息的演化性,进而提高动态文摘的性能.在 TAC2008 测试结果中排名第一,取得了较好的成绩,说明这几种模型算法结合的动态多文档文摘系统性能稳定,在动态信息处理领域具有一定的应用价值,在现在研究较少的动态文摘领域,本文的模型和算法的创新具有一定的研究意义.

此系统高效的性能可以移植到其他语种的动态多文档文摘的研究中,例如还不成熟的中文多文档文摘领域、中文动态文摘领域及中文动态信息处理领域,具有一定的研究价值和可行性.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是哈尔滨工业大学计算机科学技术学院的郑德权副教授和赵铁军教授的指导表示感谢.

References:

[1] http://duc.nist.gov/

© 中国科学院软件研究所 http://www.jos.org.cn

[2] http://www.nist.gov/tac/

[3] http://www.nist.gov/index.html

[4] http://www-nlpir.nist.gov/projects/duc/guidelines/2007.html

[5] http://www.trec.com/

[6] Allan J, Gupta R, Khandelwal V. Temporal summaries of news topics. In: Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2001). 2001. [doi: 10.1145/383952.383954]

[7] Tang XN, Yang CC. Following the social media: Aspect evolution of online discussion. In: Proc. of the Computational Linguistics and Intelligent Text Processing. Iasi: Springer-Verlag, 2010. 346−360. [doi: 10.1007/978-3-642-19656-0_41]

[8] http://www.sciencemag.org/content/253/5025/1242.abstract

[9] Mani I. Recent developments in temporal information extraction (draft). In: Nicolov N, Mitkov R, eds. Proc. of the RANLP. Inderjeet Mani: Georgetown University, 2004.

[10] Bollegala D, Okazakia N, Ishizukaa M. A machine learning approach to sentence ordering for multidocument summarization. In: Proc. of the Annual Meeting of the Association for Natural Language Processing. 2005. 482−488.

[11] Ahn D, van Rantwijk J, de Rijke M. A cascaded machine learning approach to interpretingtemporal expressions. In: Proc. of the NAACL-HLT 2007. University of Amsterdam, 2007.

[12] Piotrowski WJ, Kurmanowska Z, Antczak A, Marczak J, Górski P. Exhaled 8-isoprostane as a prognostic marker in sarcoidosis. In: Proc. of the A Short Term Follow-Up Computational Linguistics and Intelligent Text Processing. Springer-Verlag, 2010. 10−23. [doi: 10.1186/1471-2466-10-23]

[13] ACE2007 evaluation plan. 2006. http://projects.ldc.upenn.edu/ace/intro.html

[14] Song XC, Liu GQ. Multi-Document summarization method based on topic-concepts extract. Computer Engineer, 2010,36(4): 190−192 (in Chinese with English abstract).

[15] Ye N, Zhu JB, Zheng Y, Ma MY, Wang HZ, Zhang B. A dynamic programming model for text segmentation based on min-max similarity. In: Proc. of the 4th Asia Information Retrieval Symp. (AIRS 2008). 2008. 141−152. [doi: 10.1007/978-3-540-68636-1_14]

[16] Ye N, Zhu JB, Wang HZ, Ma MY, Zhang B. An improved model of Dotplotting for text segmentation. Journal of Chinese Language and Computing, 2007,17(1):27−40.

[17] Yang XX, Zhang L. Information extraction based on semantic role and concept graph. Journal of Computer Applications, 2010,30(2):411−414 (in Chinese with English abstract).

[18] Zhang J, Xu HB, Wang XL, Shen HW, Zeng YL. ICT CAS at DUC 2007. In: Proc. of the Document Understanding Conf. 2007. 231−242.

[19] Zhang J, Cheng XQ, Wu GW, Xu HB. AdaSum: An adaptive model for summarization. In: Proc. of the ACM 17th Conf. on Information and Knowledge Management (CIKM 2008). 2008. 450−463. [doi: 10.1145/1458082.1458201]

[20] Zhang J, Xu HB, Cheng XQ. Research on dynamic summarization for evolutionary Web information. Chinese Journal of Computers, 2008,31(4):696−701 (in Chinese with English abstract).

[21] Boudin F, Moreno JMT. NEO-CORTEX: A performant user-oriented multi-document summarization system. In: Proc. of the Computational Linguistics and Intelligent Text Processing. Springer-Verlag, 2010. 89−99. [doi: 10.1007/978-3-540-70939-8_49]

[22] Hovy E, Lin CY, Zhou L, Fukumoto J. Automated summarization evaluation with basic elements. In: Proc. of the Resources and Evaluation (LREC). 2006. 102−116.

[23] Hovy E, Lin CY, Zhou L. Evaluating DUC 2005 using basic elements. In: Proc. of the Document Understanding Conf. (DUC 2005). 2005. 67−78.

[24] Carbonell JG, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Information Processing & Management, 1998,31(5):675−685. [doi: 10.1145/290941.291025]

[25] Lapata M. Probabilistic text structuring: Experiments with sentence ordering. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, 2003. 545−552. [doi: 10.3115/1075096.1075165]

[26] Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. In: Proc. of the Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 412−420.

© 中国科学院软件研究所 http://www.jos.org.cn

[27] Wan X, Yang J, Xiao J. Manifold-Ranking based topic-focused multi-document summarization. In: Proc. of the IJCAI 2007. 2007. 2903−2908.

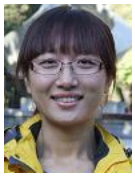
[28] Liu ML, Zheng DQ, Zhao TJ, Yu Y, Zhou JY. Text similarity cumulative model and algorithm research for dynamic multi-document summarization. Journal of Computational Information Systems, 2011,7(5):1698−1705.

附中文参考文献:

[14] 宋宣辰,刘贵全.基于主题概念抽取的多文档文摘方法.计算机工程,2010,36(4):190−192.

[17] 杨选选,张蕾.基于语义角色和概念图的信息抽取模型.计算机应用,2010,30(2):411−414.

[20] 张瑾,许洪波,程学旗.面向网络演化信息的动态文摘方法研究.计算机学报,2008,31(4):696−701.



刘美玲(1981－),女,黑龙江哈尔滨人,博士,讲师,主要研究领域为信息检索,信息处理,自动文摘技术.
E-mail: mlliu@mtlab.hit.edu.cn

任洪娥(1962－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为图像处理与模式识别,人工智能与智能控制,现代信息技术与信息安全.
E-mail: renhonge163@163.com

于洋(1977－),男,助理研究员,主要研究领域为人工智能,地理信息系统,信息检索.
E-mail: zsbyy@126.com

郑德权(1968－),男,博士,副教授,CCF 会员,主要研究领域为自然语言处理,知识工程,信息检索.
E-mail: dqzheng2007@gmail.com

赵铁军(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,人工智能.
E-mail: tjzhao@hit.edu.cn

© 中国科学院软件研究所 http://www.jos.org.cn