自动文本摘要技术综述

Summary of Automatic Text Summarization Techniques

胡 侠 林 晔 王 灿 林 立

(杭州市科学技术信息研究院 杭州 310001) (浙江大学 计算机科学与技术学院 杭州 310027)

摘 要 随着互联网上信息爆炸式的增长,如何在互联网上有效地获取所需信息成为当前情报科学领域一个迫切需要解决的问题。为了更好地浏览和吸收互联网上的海量信息,自动文本摘要技术对文档进行压缩,压缩后的表示能够覆盖原文的所有主题且不重复。文章对目前单文档摘要和多文档摘要领域的一些最相关技术和方法做一个较为全面的综述性介绍,对该领域当前的一些最新发展趋势,如基于图排序的摘要方法也进行了简要的探讨。

关键词 自动摘要 文档抽取 机器学习

中图分类号 TP391 文献标识码 A 文章编号 1002-1965(2010)08-0144-04

1 背 景

随着 Internet的飞速发展,人们越来越多地依赖于万维网来获取所需要的信息。如何更加有效地浏览和查阅万维网上的海量信息成了当前情报科学领域的研究热点。自动文本摘要技术对文档信息进行压缩表示,更好地帮助用户浏览和吸收万维网上的海量信息。在万维网用户普遍面临信息过载问题的今天,自动文本摘要技术无疑能够有效地降低用户的信息负载,帮我们更好地从万维网获取各类科技情报信息。近年来,自动文本摘要技术在科技情报领域的应用不断扩展,有效提高了科技工作者浏览和处理信息的效率,是当前信息检索领域的研究热点之一。

2 研究现状

自动文本摘要技术从 20世纪 50年代开始兴起,最初是以统计学为支撑,依靠文章中的词频、位置等信息为文章生成摘要,主要适用于格式较为规范的技术文档。从 90年代开始,随着机器学习技术在自然语言处理中的应用,自动文本摘要技术中开始融入人工智能的元素。针对新闻、学术论文等主题明确、结构清晰的文档,一些自动摘要技术[1-2]使用贝叶斯方法和隐马尔可夫模型抽取文档中的重要句子组成摘要。到了 21世纪,自动文本摘要技术开始广泛应用于网页文档。针对网页文档结构较为松散、主题较多的特点,网页文档摘要领域出现了一些较新的自动摘要技术,比如基于图排序的摘要方法等。

我们可以根据自动文本摘要技术本身的特点对其进行分类。根据摘要的主题聚焦性,自动文本摘要又可分为普适摘要和查询相关的摘要。其中,普适摘要会尽量覆盖文章中的所有主题并将冗余最小化;而查询相关的摘要则是抽取文章中和查询词紧密相关的内容。所产生的摘要从形式上可以分为文摘(extract)和摘要(abstract),文摘通过抽取原文中的重要句子所组成,而摘要则对相关语义信息用新的句子进行描述。目前,大多数的摘要方法都是基于文摘的方法。

根据摘要所覆盖的文档数量,自动文本摘要可以分为单文档摘要与多文档摘要。单文档摘要技术为单个文档生成摘要,而多文档摘要技术则为多个主题类似的文档产生摘要。本文将在接下来的篇幅中对单文档摘要技术、多文档摘要技术以及新兴的网页摘要技术做一个概述性的介绍。

2.1 单文档自动摘要技术

单文档自动摘要技术针对单个文档,对其中的内容进行抽取,并针对用户或者应用需求,将文中最重要的内容以压缩的形式呈现给用户。常见的单文档摘要技术包括基于特征的方法、基于词汇链的方法和基于图排序的方法。

2.1.1 基于特征的方法。

文档摘要中常用的文章特征包括词频、特定段落(如首末段)、段落的特定句子(如首末句)等。Luhn在 1958年发表的论文[3]指出,频繁出现的单词与文章主题有比较大的关联,因此可以根据各单词出现的频率给文中的句子打分,以得

收稿日期: 2010-04-02 修回日期: 2010-06-11

作者简介: 胡 侠(1974-),女,硕士,助理研究员,研究方向为情报理论、方法及应用;林 晔(1962-),男,研究员,研究方向为情报理论、方法及应用;王 灿(1974-),男,博士,工程师,研究方向为数据挖掘;林 立(1985-),男,硕士,研究方向为信息检索、网络系统研发。

分最高的几个句子组成文章的摘要。有趣的是, 后来的评估表明[4], 这个看似最简单的方法, 准确率却比后来不少复杂的方法要高。

Baxendale等人通过从句子位置特征入手, 通过计算文章中段落首末句出现主题句的概率, 选取得分最高的若干句子生成摘要[5]。Edmundson利用线索词(cue words)、标题词、句子位置以及关键词频等 3个因素, 计算每个句子的权重, 得分最高的几个句子作为摘要[6]。

到了 20世纪 90年代, 随着机器学习在自然语言处理领域应用的兴起, 自动摘要技术中也逐渐开始出现一些基于机器学习的方法。在 Edmundson的研究基础上[6], Kupiec在 1995年提出一种新的方法[1], 通过朴素贝叶斯分类模型去判定文章里的每个句子是否应该抽取为摘要。在 Kupiec的方法中, 假设 s是某一句子, S是组成摘要的句子集合, $F_1, \ldots, F_k$为文章的 k个特征, 假设这 k个特征相互独立, 则有以下公式:

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) \cdot P(s \in S)}{\prod_{j=1}^{k} P(F_j)} \quad (1)$$

通过公式(1)计算出每个句子成为文章摘要的概率, 最后得分最高的几个句子抽取出来作为文章的摘要。Aone等人在 1999年开发出一个基于贝叶斯分类模型的系统 DimSim[7], 在这个系统中他们采用了更多的文章特征来计算句子的摘要概率, 如词组频率(item frequency)以及倒文档频率(inverse document frequency)等。他们在该系统中使用了词组别名的匹配方法, 例如把 IBM与 International Business Machines等同起来, 从而改善了摘要的质量。

通过对句子位置进行深入的分析, Lin与 Hovy根据每个句子的位置加权计算句子的分值[8]。他们将该方法应用到了针对新闻类文章摘要的 TIPSTER系统, 并在该系统中将加权规则针对一般文章也做了优化。但由于不同文章的逻辑结构往往不同, 这个方法只在特定的领域才会有较好的摘要效果。

Lin等人在 1999年提出另一种摘要方法[9]。在这种方法中, 他们假设文章中用于摘要抽取的各种特征是相互关联的, 并使用了决策树(decision trees), 而不是贝叶斯分类模型对句子打分, 抽取得分最高的部分句子作为文章摘要。

另外, 在 Osborne等人提出的基于数线性模型(Log-Linear Models)的摘要方法中[10], 他们注意到了各种特征间的关联性, 并通过实验证明了这种模型比朴素贝叶斯模型的提取效果要好。该模型可以用下面的公式表示:

$$P(l \mid s) = \frac{\exp(\sum_i \lambda_i f_i(l, s))}{\sum_l \exp(\sum_i \lambda_i f_i(l, s))} \quad (2)$$

其中 l是标签(在该模型里存在两种标签: 该句子被抽取为摘要或不被抽取为摘要), s是要标注的某个项, 为对应特征的权重。

Conroy与 O'leary在 2001年提出一种使用隐马尔可夫模型(hidden Markov model)的摘要方法[2]。该方法也使用了一些文章的特征来确定句子的分值, 如句子位置、句内词数以及句内词语与文章词语的相似度等。

2.1.2　基于词汇链的方法。基于词汇链的方法主要通过对文章内容进行自然语言分析生成摘要。这类方法中, 有代表性的方法是 Miller在 1995年提出的[11]。该方法通过分析生成词汇链(lexical chain)来做摘要提取, 主要分为 3个步骤: a选择候选词的集合; b根据与词汇链里成员的相关程度, 为每个候选词选择词汇链; c如果发现候选词与某词汇链相关度高, 则把候选词加入词汇链内。

最后该方法根据长度与一致性给每个链打分, 并使用一些启发式方法挑选部分词汇链生成摘要。在此基础上, Ono等人在 1994年提出了结合修辞结构的应用[12]; Marcu则更进一步地提出了修辞学理论[13-14]。Marcu把文章中的文字段分为两类: 中心段与随从段, 并把这些文字段建立成树状关系并以此生成摘要。

2.1.3　基于图排序的方法。基于图排序的文本摘要方法的一般思想是把文章分解为若干单元(句子或段落等), 每个单元对应一个图的顶点, 单元间的关系作为边, 最后通过图排序的算法(如 PageRank, manifold ranking等)得出各顶点的得分, 并在此基础上生成文本摘要。

在以句图结构表示文档的基础上, Mihalcea等人使用了 PageRank算法来提取出关键的句子生成文档摘要[15]。在该方法中, 他们把每个句子作为图的顶点, 句子间的相似度作为顶点间的边。句子间的相似度由句子内容的重叠程度决定, 通过两个句子间的共同单词数量计算而得。为了避免长句子分数过高的情况, 他们把得出的数值与句子长度相除。只有在两个句子间的相似度大于零时, 它们对应的顶点才会有边相连。文章对应图的生成有 3种建模方法: 无向加权图; 有向加权图, 边的方向顺着文章句子顺序, 边的权重为两句子间的相似度; 与第二种方法方向相反的有向加权图。最后, 他们使用了 HITS, PageRank与无向图的联通性等方法进行了试验, 最后得出每个句子对应的分数, 由得分最高的句子组成文章的摘要。

耿焕同等人则利用句子间的共同词, 提出了一种

基于词共现图的文档自动摘要算法[16],通过词共现图形成的主题信息以及不同主题间的连接特征信息自动地提取文档摘要。

2.2 多文档自动摘要技术 多文档自动摘要的目的是为包含多份文档的文档集合生成一份能概括这些文档主要内容的摘要。相对单文档自动摘要,多文档自动摘要除了要剔除多份文档中的冗余内容外,还要能够识别不同文档中的独特内容,使得生成的摘要能够尽量的简洁完整。

多文档自动摘要的研究从20世纪90年代开始兴起,尽管目前还没有非常满意的解决方案,但不少人员组织一直在做各种尝试,如Google公司的Google News http://news.google.com,哥伦比亚大学的Columbia NewsBlaster http://newsblaster.cs.columbia.edu/等。该领域一个较早的工作来自于哥伦比亚大学的自然语言处理小组,他们在1995年开发出SUMMONS系统(SUMMarizing Online NewS),并在新闻领域的多文档摘要取得不错效果[17]。有些多文档摘要方法通过聚类(clustering)方法来识别文档集合中的共同主题,并从每个聚类中摘取句子组成摘要[18-19],或者是从各聚类中生成一个重新组合过的句子[20]。还有些方法使用最大边缘相关(maximal marginal relevance)理论评估每个段落,并使用重要的段落组成最终摘要[21]。最早的多文档摘要技术只能处理同一语言的文档集合,但后来的一些研究把该技术拓展到多语言环境[22]。

多文档自动摘要领域一个比较有代表性的方法是Erkan等人提出的LexRank方法[23]。与Mihalcea在单文档摘要领域的工作[15]类似,LexRank方法也通过句子间的相似性来为多文档构建句图。不同的是,LexRank方法使用到词频(term frequency,即tf)与倒排文档频率(inverse document frequency,即idf)来衡量句子间的相似性。tf指一个单词在某文档中出现的次数,idf的计算公式如下:

$$idf_i = \log\left(\frac{N}{n_i}\right) \quad (3)$$

其中N代表集合中的文档数量,n_i表示单词i在n_i个文档出现。Erkan等人把文档中的句子构建成一个N维的向量,假设$tf_{w,s}$表示词w在句子s中的出现次数,则句子x与句子y的相似度计算公式为:

$$similarity(x,y) = \frac{\sum_{w\in x,y} tf_{w,x}\, tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i\in x}(tf_{x_i,x}\, idf_{x_i})^2} \times \sqrt{\sum_{y_i\in y}(tf_{y_i,y}\, idf_{y_i})^2}} \quad (4)$$

Carbonell与Goldstein提出了主题驱动式的多文档摘要(Topic-driven Summarization)方法[24],该方法使用最大边缘相关度模型去除多文档内的冗余内容并选择合适的段落来组成摘要。刘德荣等人提出了一种基于主题概念的多文档自动摘要方法[25],通过对文档主题概念的关联分析判断多文档间的相关度,并利用HOWNET(一个描述有关概念及其属性之间的关系的知识库)来计算文献主题概念的内聚度实现多文档的自动摘要。另外,Mani与Bloedorn使用基于图的方法[26]来发现不同文档中的相似内容和相异内容,并通过对相异内容评分排序,抽取得分最高的部分组成多文档摘要。

2.3 网页文档自动力摘要技术 相较于传统的文档,网页文档有着结构较为松散、主题多样化等特点。同时,除了文档文本的内容,网页中往往还会有一些额外的信息可以用于文档摘要,比如网页上的评论、标签[27]等。这些额外信息往往与文章主题高度相关,同时也是用户关注的焦点。利用这些信息,可以使产生的网页摘要有效聚焦于用户所普遍感兴趣的主题。

Meishar等人[28]把网页里面的评论关系区分为3种:主题、引用与提及,并把它们之间的关系建模成3种图,并使用基于图和基于张量(tensor-based)方法对每个评论打分以评估其重要性,最后使用基于特征方法或统一文档方法(uniform-document approach)从文档中提取出句子组成网页文档摘要。

Sun等人[29]认为对某一网页进行操作的用户对网页内容应该是有所理解的,比如用户点击网页链接时,往往对链接所指向的页面内容会有一个初步判断。基于这个设想,他们提出了一种结合网页链接点击生成网页内容摘要的方法。另外,马慧芳等人提出了一种基于文本关系图的网页文档摘要技术[30],利用搜索引擎的返回结果,为多个网页文档自动产生摘要,以提高搜索引擎使用效率。

Jaehui Park与Tomohiro Fukuhara[31]通过社群书签(Social Bookmarks,比如del.icio.us,Digg,YouTube与Amazon.com等)里面的评论与标签入手去生成文章摘要。他们开发了SSNote系统来对分析del.icio.us的评论与标签,并提取出摘要。

3 总 结

互联网的迅速发展给我们提供了海量信息,同时也使信息的有效获取日益成为当前情报科学领域一个迫切需要解决的问题。自动文本摘要技术将冗长的文档内容压缩成较为简短的几段话,从而加速信息理解和吸收,有效解决信息过载问题。本文对当前自动文本摘要技术中的主流方法进行了综述,介绍了自动文本摘要领域的3类主要技术:单文档摘要、多文档摘要以及新兴的网页摘要技术。

自动文本摘要技术的研究已经持续了60多年,所

应用的摘要对象也从专业文献到新闻、电子邮件延伸到了万维网网页。如何在摘要提取中充分利用各类不同文档的结构和内容特征, 将是提升自动文本摘要效果的关键; 而机器学习技术的兴起, 则为这些特征的有效利用提供了可能。在自动文本摘要中应用各类机器学习算法, 以达到更佳的摘要效果, 将是该领域以后的主要研究方向。

参 考 文 献

[1] Kupiec J, Pedersen J, Chen F. A Trainable Document Summarizer[C]. ACM SIGIR, New York, USA, 1995

[2] Conroy J M, O'leary D P. Text Summarization Via Hidden Markov Models[C]. ACM SIGIR, New Orleans, Louisiana, USA, 2001

[3] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research Development, 1958, 2(2): 159

[4] Text Summarization[EB/OL]. http://www.summarization.com/sigirtutorial2001.ppt

[5] Baxendale P. Machine-made Index for Technical Literature—an Experiment[J]. IBM Journal of Research Development, 1958, 2(4): 354

[6] Edmundson H P. New Methods in Automatic Extracting[J]. Journal of the ACM, 1969, 16(2): 264

[7] C Aone M E, Okurowski J Gorlinsky and B Larsen. A Trainable Summarizer With Knowledge Acquired from Robust NLP Techniques// J Mani and M Maybury (eds). Advances in Automated Text Summarization. ed. MIT Press, 1999. 71

[8] Lin C Y, Hovy E H. Identifying Topics by Position[C]. The Applied Natural Language Processing Conference, Washington DC, USA, 1997

[9] Lin C Y. Training a Selection Function for Extraction[C]. the Eighth ACM Conference on Information And Knowledge Management, Kansas City, Missouri, USA, 1999

[10] Osborne M. Using Maximum Entropy for Sentence Extraction[C]. ACL'02 Workshop on Automatic Summarization, Philadelphia, USA, 2002

[11] Miller G A. Wordnet: a Lexical Database for English, Commun[J]. Communications of the ACM, 1995, 38(11): 39

[12] Ono K, Sumita K, Miike S. Abstract Generation Based on Rhetorical Structure Extraction[C]. International Conference on Computational Linguistics, Kyoto, Japan, 1994

[13] Marcu D. Improving Summarization Through Rhetorical Parsing tuning[C]. The Sixth Workshop on Very Large Corpora, Montreal, Canada, 1998

[14] Marcu D C. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts[C]. the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, 1997

[15] Rada Mihalcea. Graph-based Ranking Algorithms for Sentence extraction, Applied to Text summarization[C]. the ACL 2004 on Interactive Poster and Demonstration Sessions, Barcelona, Spain, 2004

[16] 耿焕同, 蔡庆生, 赵　鹏等. 一种基于词共现图的文档自动摘要研究[J]. 情报学报, 2005, 24(6): 652

[17] McKeown K R, Radev D R. Generating Summaries of Multiple News articles[C]. SIGIR'95, Seattle Washington, 1995

[18] McKeown K, Klavans J, Hatzivassiloglou V, et al. Towards Multidocument Summarization by Reformulation: Progress and prospects[C]. the 16th National Conference on Artificial Intelligence (AAAI'99), Orlando, FA, USA, 1999

[19] Radev D R, Jing H, Budzikowska M. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies[C]. NAACL-ANLP 2000 Workshop on Automatic Summarization, Morristown, NJ, USA, 2000

[20] Barzilay R, McKeown K, Elhadad M. Information fusion in The Context of Multi-document Summarization[C]. of the 37th Conference on Association for Computational Linguistics (ACL'99), College Park, Maryland, MD, USA, 1999

[21] Carbonell J, Goldstein J. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries[C]. SIGIR'1998, New York, NY, USA, 1998

[22] Evans D K. Similarity-based Multilingual Multi-document Summarization[R]. Technical Report CUCS-014-05, Columbia University

[23] Erkan G, Radev D. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization[J]. Journal of Artificial Intelligence Research, 2004(22): 457

[24] Carbonell J, Goldstein J. The use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries[C]. SIGIR'98, New York, NY, USA, 1998

[25] 刘德荣, 王永成, 刘传汉. 基于主题概念的多文档自动摘要研究[J]. 情报学报, 2005, 24(1): 69

[26] Mani, I Bloedorn E. Multi-document Summarization by Graph Search and Matching[C]. the Fourteenth National Conference on Artificial Intelligence, Providence, Menlo Park, California, USA, 1997

[27] Junyan Zhu, Can Wang, Xiaofei He, et al. Tag-Oriented Document Summarization[C]. the 18th International Conference on World Wide Web, Madrid, Spain, 2009

[28] Meishan Hu, Aixin Sun, Ee-Peng Lim. Comments-oriented Document Summarization: Understanding Documents With Readers' feedback[C]. the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, Singapore, 2008

[29] J T Sun, D Shen, H-J Zeng, et al. Web-page Summarization Using Clickthrough data[C]. SIGIR'05, Salvador, Brazil, 2005

[30] 马慧芳, 祁云平, 杨小东. 一种基于文本关系图的多文档自动摘要技术[J]. 情报杂志, 2007(3): 67

[31] Jaehui Park, Tomohiro Fukuhara, Lkki Ohmukai, Web Content Summarization Using Social bookmarks: a New Approach for Social Summarization[C]. the 10th ACM Workshop on Web Information and Data Management, Napa Valley, California, USA, 2008

(责编: 王平军)