文章编号:1001 - 9081(2006)11 - 2626 - 02

多文档文本摘要的一种改进 HITS 算法

黄丽雯,钱 微
(重庆工学院 电子信息与自动化系,重庆 400050)
(cqhlw@163.com)

摘 要:提出了一种对 HITS 算法进行改进的新方法,本方法将文档内容与一些启发信息如“短语”,“句子长度”和“首句优先”等结合,用于发现多文档子主题,并且将文档子主题特征转换成图节点进行排序。通过对 DUC 2004 数据的实验,结果显示本方法是一种有效的多文本摘要方法。

关键词:图节点;多文档;文本摘要;子主题
中图分类号: TP391.1 **文献标识码**:A

An improved HITS approach for multi-document text summarization

HUANG Li-wen, QIAN Wei
(*Department of Electronic Information and Automation, Chongqing Institute of Technology, Chongqing* 400050, *China*)

Abstract: A new approach for improving Hyperlink-Induced Topic Search (HITS) algorithm was proposed. The approach combined the text content with some cues such as cue phrase, sentence length and first sentence, so as to explore the sub-topics in the multi-documents by bringing the features of these sub-topics into graph-based sentence for ranking. Experiment on DUC 2004 data show that the new approach is an effective technique in multi-document generic text summarization.

Key words: graph node; multi-document; text summarization; sub-topics

0 引言

互联网的快速增长并由此产生的大量信息使得有效访问所需信息变得越来越困难,自动文摘并向终端用户提供压缩后意义仍然完整的结果能有效地解决这个问题。文本摘要的基本过程是基于现有文档自动创立一个压缩版本,然后为用户提供有用的信息。

选择原始文档的句子子集来提取摘要较为常用,此过程从与主题有关的句子集合中提取最重要的句子。在基于子主题的摘要提取过程中评估词汇的重要性是依据文档质心,包含较多主题词汇的句子被认为是核心句子。其他方法还包括监督方法,根据初始摘要和给予的文档中句子间的相似性提取摘要。文献[1]讨论了有关自动句子摘要提取的有效性和局限性,强调利用准确工具作句子提取应作为自动提取系统的整体部分。

近年来业界已提出了一些基于图节点排序的文本摘要方法。LexPageRank[2]是一种基于特征向量中心性概念的句子重要性计算方法,在此模型中,语句相似性是以向量空间模型为基础的,如果两个句子间相似性超过特定值,相应的边将被添入图中作为连接图中节点的边。PageRank 算法的改进算法被应用于语句重要性排序。Rada 调查了基于图的排序算法,对无监督的文本摘要作了评价,此种无监督算法所获得的结果与以前方法相比有一定优势。

本文提出在 HITS 算法框架[3]下的一种新方法,该方法将文档内容与一些启发信息如“短语”,“语句长度”和“首句”结合,并用于发现多文档子主题,将子主题特征转换成图节点排序。

1 基于启发信息的 HITS 文本摘要提取法

基于图的句子排序算法本质上是通过一定算法来衡量图节点的重要性,由于启发规则是重要的文本摘要可利用信息,本文在 HITS 框架下将启发规则与内容特点结合起来对句子进行排序。本文提出的改进算法中利用了 3 个启发规则:首句、短语启发信息和句子长度启发信息。

首句 文档中段落的第一句是比较可靠的摘要信息,因为在多数文章中第一句常常简要描述文章的内容。

短语启发信息 一些特殊短语比如“总而言之”或“总之”等常跟有重要信息,因此,含有更多的短语启发信息被认为比没有短语启发信息的句子更加重要。

句子长度 短句信息量小,长句常比短句包含信息丰富。

本算法首先用句子聚类的方法发现多文档的子主题并提取不同子主题下的特性词(或短语)集合;然后,所有的特性词和短语启发信息被用来作为图中 Authority 节点,所有句子被视为 HUB 节点,如果句子包含了 Authority 里的词汇,在 Authority 词和 HUB 句子间将有一条边。每个节点的初始权值包含内容和启发信息诸如短语启发信息和首句,通过 HITS 算法相互加强机制,能够在多文档中给语句进行排序。基于启发信息的 HITS 算法的基本假设是好 Authority 词汇(或短语)是指向很多好 HUB 句子的内容,一个好的 HUB 句被很多好的 Authority 词指向。最后,我们采用 Markov 模型确定子主题顺序,最终摘要将按用户需求将句子排序输出。

1.1 子主题提取

多个文档可被看作相互关联的句子集合,其中一些句子比较类似,而另外一些则差异较大,假定那些与很多其他句子相

收稿日期:2006 - 05 - 08;**修订日期**:2006 - 08 - 23

作者简介:黄丽雯(1967 -),女,江苏南京人,副教授,硕士,主要研究方向:信息处理、远程测试及处理; 钱微(1965 -),男,陕西户县人,高级工程师,硕士,主要研究方向:工业自动化、实验室管理

似的句子集合为表达文档主题的某一特定的子主题，并以此将文档的句子集合划分为不同子集合来获得初始的“子主题”集合[4]，通过 K 最相邻(K-Nearest Neighbour，KNN)聚类获得 K 个不同子主题(利用向量空间模型计算句子相似性)。

通过聚类不仅提取到作为 HITS 算法节点的信息还在句子选择过程中为最终形成摘要决定了子主题顺序。

TF * IDF 算法用于提取子主题特征词汇，TF 表示词汇出现频率，IDF 表示在不同子主题中出现的次数。得分最高的 k 个词汇被提取出来作为子主题特性集合。

1.2　计算 Authority 节点和 HUB 节点给句子排序

本文提出的在 HITS 框架下将启发规则和内容特性结合起来给句子进行排序的方法是基于图排序的算法。图由两类节点构成:Authority 和 HUB。Authority 节点是一些内容或启发信息词汇(或短语)，HUB 节点是多文档中的语句。如果句子包含 Authority 中的词汇，从 Authority 有边指向 HUB 语句。Authority 和 HUB 体现了交互加强关系：好的词语(或短语)是一个指向很多好 HUB 句子的文本内容，好的 HUB 句子是被很多好的 Authority 词指向的节点。

通过迭代算法利用 Authority 和 HUB 关系更新节点权值。对于 Authority 中每个节点，赋予非负向量 $\boldsymbol{x}(h_i)$，而对于 HUB 节点赋予非负向量 $\boldsymbol{y}(a_j)$。初始 Authority 节点权值设定遵循以下方法:短语是较为重要的启发规则，初始 HUB 节点是由“首句优先”的启发规则决定的。对于两种类型的向量进行归一化操作，使得它们的平方和为 1：$\sum_{v\in HUB}\boldsymbol{x}^2(h_v)=1$，$\sum_{v\in Authority}\boldsymbol{y}^2(a_v)=1$，具有较大 $\boldsymbol{x}$ 值和 $\boldsymbol{y}$ 值的节点分别为较好 Authority 和 HUB。

HITS 算法体现了 Authority 与 HUB 间的相互加强的关系:若 Authority 节点 h_i 指向很多“$\boldsymbol{y}$ 值大的语句，那么它将接收一个大 $\boldsymbol{x}$ 值；而若很多 $\boldsymbol{x}$” 值大的词汇指向 HUB 节点 a_i，该节点将接收大 $\boldsymbol{y}$ 值。这引出了权值计算的两种操作(以 I 和 O 表示)被赋予向量$\{\boldsymbol{x}(h_i)\}$，$\{\boldsymbol{y}(a_j)\}$，I 操作更新 $\boldsymbol{x}$ 权值：

$$\boldsymbol{x}(h_i)\leftarrow\sum_{a_j:\,(h_i,\,a_j)\in E}\boldsymbol{y}(a_j)\tag{1}$$

O 操作更新 $\boldsymbol{y}$ 权值：

$$\boldsymbol{y}(a_j)\leftarrow\sum_{h_i:\,(h_i,\,a_j)\in E}\boldsymbol{x}(h_i)\tag{2}$$

因此 I 和 O 操作是 Authority/HUB 相互加强的基本手段。

为达到最优效果，以 I 和 O 操作交互运用的方式计算看是否达到一个稳定值。计算 Authority 和 HUB 权值的算法 HITS(G,k)的算法流程如下：

输入 G：图节点和边的集合；

输入 k:迭代次数；

Z :初始 Authority 向量，如果 Authority 词是启发信息词，初始权值设为 2，否则设为 1；

W :初始 HUB 向量权值，如果 HUB 语句是段落第一语句，初始权值设为 2，否则设为 1；

1) 设 $x_0=z$，$y_0=w$；$i=1$；

2) 比较 i，k 大小，如果 $i\leqslant k$，执行 3)，否则执行 5)；

3) 应用 I 操作，得到新的 $\boldsymbol{x}$ 权值 x'_i；应用 O 操作，得到新的 $\boldsymbol{y}$ 权值 y'_i；

4) 规格化 x'_i，得到 x_i；规格化 y'_i，得到 y_i；$i=i+1$；

5) 输出 x_k 和 y_k。

从算法 HITS(G,k)可知“句子长度”因素已经被隐含包括在算法中：句子越长，权值越高。执行算法后，得到最终 Authority 权值向量 x_k 和 HUB 向量 y_k。HUB 向量中的权值可被看作文档所包含的不同语句的重要性得分。

1.3　形成摘要

多文档摘要最后的过程是以一定的顺序组织所选句子，以使读者能够理解自动生成的摘要，本文根据子主题顺序选择语句，每个子主题代表了提及的子主题中的所有语句。假设多个文档有 m 个不同的子主题 $T=\{T_1,T_2,T_3,\cdots,T_m\}$，那么面临的问题是怎样安排这些子主题的顺序。

假设文档 $d=\{S_1,S_2,\cdots,S_j,\cdots,S_n\}$ 包含 n 个语句和 l 个子主题 $T_i=\{s'_1,s'_2,\cdots,s'_j,\cdots,s'_l\}$，假设交集是 $I=d\cap T_i=\{s''_1,s''_2,\cdots,s''_j,\cdots,s''_l\}$，用 I 和 S_i 的最大得分选择语句 S_i 作为文档 d 的子主题 T_i 的代表语句。重复此过程，获得不同子主题的 m 个代表语句，所有这些语句属于文档 d，再将文档 d 中子主题的顺序确定为此文档的子主题顺序。

多文档子主题顺序利用 Markov 模型计算，其中的状态代表不同子主题，而状态转换代表在多文档中的信息表达顺序。

假设多文档集合包含了 K 个文档，对任何单文档，按以上方法决定文档子主题顺序。比如 D 文档子主题顺序如下：

$(D)=ST_1\,T_2\cdots T_nE$

这里 S 指开始状态，而 E 指结束状态。

所以有类似的 K 个子主题转移序列，利用二元语言模型计算其状态转移概率如下：

$$P(T_j\mid T_i)=\frac{P(T_iT_j)}{P(T_i)}\tag{3}$$

以计算下列概率的最大值来决定多文档子主题顺序：

$$\operatorname{argmax}\,P(ST_{k1}T_{k2}\cdots T_{kn}E)=$$
$$P(S)P(T_{k1}\mid S)P(T_{k2}\mid T_{k1})\cdots P(E\mid T_{kn})\tag{4}$$

$ST_{k1}T_{k2}\cdots T_{kn}E$ 指任何从开始到结束的状态转移过程，同时 $P(S)=1$。

在子主题顺序计算好后，挑选最高得分的句子作为摘要。一般说来，最终的摘要由用户决定长度，这样就能按照得分排列所选句子，然后从高分到低分输出语句。一旦所需长度达到时，就终止此过程。

2　实验结果

最后，本文采用了 DUC 2004 的英文摘要数据作为实验数据，评价工具使用 ROUGE 工具(以 N 元统计为基础的方法，与人工评价高度相关[2])。将自动摘要长度确定为 665 byte，且与人工摘要对照计算 ROUGE 得分。在实验结果中列出了三种 ROUGE 计算方法：ROUGE-1 (一元为基础)，ROUGE-2 (二元为基础)，和 ROUGE-LCS (以最长的共同顺序为基础)。

表 1　实验结果

方法	ROUGE-1	ROUGE-2	ROUGE-LCS
随意抽取	0.3149	0.0573	0.1059
位置抽取	0.3528	0.0855	0.1293
本方法	0.3761	0.0908	0.1326

(下转第 2630 页)

有效地计算 $P(t_i \mid Q)$ 的值,可以简化为如下两种情况:

情况 1 对于任意术语 T_i,若 $Pa(T_i)$ 中的所有术语都未在查询 Q 中出现,(5) 式的最终结果为:$P(t_i \mid Q) = 1/M$。

情况 2 对于任意术语 T_i,若 $Pa(T_i)$ 中的术语在查询 Q 中出现,则分下列四种情况讨论:

① $Pa(T_i)$ 中的所有术语都在查询 Q 中出现,这种情况比较少见,(5) 式的最终结果为 $P(t_i \mid Q) = 1.0$;

② 只有 T'_i 在查询 Q 中出现,(5) 式转化为 $P(t_i \mid Q) = \frac{1-\beta}{M} + \beta$;

③ 只有 T_i 的部分或全部同义词在查询 Q 中出现,(5) 式转化为 $P(t_i \mid Q) = \frac{1-\beta}{\mid Pa(T_i) \mid - 1} \sum_{T'_j \in Pa(T_i),\ i \neq j} P(t'_j \mid Q) + \frac{\beta}{M}$;

④ T'_i 和 T_i 的部分同义词在查询 Q 中出现,这种情况也比较少见,(5) 式转化为 $P(t_i \mid Q) = \frac{1-\beta}{\mid Pa(T_i) \mid - 1} \sum_{T'_j \in Pa(T_i),\ i \neq j} P(t'_j \mid Q) + \beta$。

2) 基于以上推理,计算文档 D_j 的最终后验概率:

$$P(d_j \mid Q) = \sum_{T_j \in Pa(D_j)} w_{ij} P(t_i \mid Q) \qquad (6)$$

最后,文档以概率递减的顺序呈现给用户,这样就完成了整个信息检索过程。

4 实验与分析

表 1 SBN 与 EBN-β 的 Recall-Precision 对照

Recall	Precision					
	SBN (EBN-1.0)	EBN-0.9	EBN-0.8	EBN-0.7	EBN-0.6	EBN-0.5
0.1	0.7576	0.7687	0.8223	0.8871	0.9190	0.9504
0.2	0.6807	0.7049	0.7921	0.8532	0.8724	0.9375
0.3	0.6704	0.6599	0.7506	0.7873	0.8042	0.8372
0.4	0.6261	0.6572	0.7145	0.7523	0.7739	0.7578
0.5	0.5664	0.6481	0.7021	0.7188	0.7199	0.7566
0.6	0.5275	0.6290	0.6743	0.6770	0.6695	0.7232
0.7	0.4991	0.6027	0.6426	0.6479	0.6240	0.7007
0.8	0.4479	0.5936	0.6243	0.6208	0.5876	0.6315
0.9	0.3872	0.5441	0.5971	0.5864	0.5619	0.5820
1.0	0.2220	0.3767	0.4502	0.4648	0.4284	0.4177

实验所用文档来源于中国学术期刊网全文数据库。从该数据库共下载 701 篇文档作为文档测试集合,经处理后这些文档被 1083 个代表文档主要内容特征的术语索引,针对这些文档共构造 18 个查询。为了准确比较简单模型和扩展模型的性能,参数 β 取 6 个不同的值(0.5,0.6,0.7,0.8,0.9,1.0)进行实验,分别比较它在 10 个标准的查全率(Recall)值所对应的平均查准率(Precision)值。实验结果如表 1 所示。从实验数据可以看出:扩展模型(EBN-β) 的检索性能明显优于简单模型(SBN),而且通过调节参数 β 的取值改变扩展模型中术语间的强度关系可以获得更理想的检索效果。$\beta = 0.5$ 时,扩展模型的检索效果最佳,但是对于只有一个同义词的术语而言,缺乏辨别同义词的能力;对于 β 取其他值的情况,如 $\beta = 0.6$ 和 $\beta = 0.7$,检索效果比较理想;$\beta = 1.0$ 时,扩展模型等价于简单模型。

5 结语

文章利用同义词表示术语间关系的拓扑结构,提出一个扩展的贝叶斯网络检索模型,并通过实验将新模型和原模型的检索性能进行分析与比较。结果表明:新模型可以在不偏离用户检索目标的前提下,扩大相关信息的检索,尤其是检索非专业类文档,这主要是因为本实验所用的同义词识别工具——《同义词词林(扩展版)》,目前收录的词汇大部分是一般意义上的同义词而非专业领域的同义词,随着同义词识别技术的不断完善以及各种义类词典所收录的词汇不断扩充,所提模型会具有更好的应用价值。

致谢:本实验所用的同义词识别工具——《同义词词林(扩展版)》,由哈尔滨工业大学信息检索实验室刘挺教授提供,在此表示感谢!

参考文献:

[1] 殴洁,林守勋. 基于贝叶斯网络模型的信息检索[J]. 微电子学与计算机, 2003, 20(5): 83 - 87.

[2] DE CAMPOS LM, FERNANDEZ-LUNA JM, HUETE JF. Clustering terms in the Bayesian network retrieval model: a new approach with two term-layers [J]. Applied Soft Computing, 2004, 4(2): 149 - 158.

[3] DE CAMPOS LM, FERNANDEZ-LUNA JM, HUETE JF. Bayesian networks and information retrieval: an introduction to the special issue[J]. Information Processing and Management, 2004, 40(5): 727 - 733.

[4] 陆勇,侯汉青. 用于信息检索的同义词自动识别及其进展[J]. 南京农业大学学报(社会科学版), 2004, 4(3): 87 - 93.

[5] ACID S, DE CAMPOS LM, FERNANDEZ-LUNA JM, *et al.* An information retrieval model based on simple Bayesian networks[J]. International Journal of Intelligent Systems, 2003, 18(2): 251 - 265.

[6] DE CAMPOS LM, FERNANDEZ-LUNA JM, HUETE JF. The BNR model: foundations and performance of a Bayesian network-based retrieval model[J]. International Journal of Approximate Reasoning, 2003, 34(2/3): 265 - 285.

(上接第 2627 页)

为便于比较,设计了 2 个基准测试:1)位置抽取方法。提取每篇文章第一语句产生摘要。2)随机选取语句。这里的“随机”指的是随机从句子集合里面挑选语句的办法。经过 5 次随意选择后,挑选中值作为最终结果,实验结果见表 1。

结果证明对于多文档文本摘要,在 HITS 框架下结合启发规则和内容特征是一种有效的摘要算法。另一个方面,Authority 中的词汇能作为关键词来阐明一些文档中的主题。

参考文献:

[1] LIN CY, HOVY EH. The potential and limitations of sentence extraction for summarization[A]. Proceedings of the HLT/NAACL Workshop on Automatic Summarization[C]. Edmonton, Canada, 2003.

[2] ERKAN G, RADEV D. LexPageRank: Prestige in Multi-Document Text Summarization[A]. Proceedings of EMNLP 2004[C]. Barcelona, Spain, 2004.

[3] KLEINBERG JM. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604 - 632.

[4] WU J, KHUDANPUR S. Building a topic-dependent maximum entropy language model for very large corpora[A]. Proceedings of ICASSP[C], 2002, 1. 777 - 780.