基于 DSC 的多文本自动摘要①

李成果

(中国科学院软件研究所, 北京 100190)
(中国科学院大学, 北京 100190)

摘 要: 多文本摘要的目标是对给定的查询和多篇文本(文本集), 创建一个简洁明了的摘要, 要求该摘要能够表达这些文本的关键内容, 同时和给定的查询相关. 一个给定的文本集通常包含一些主题, 而且每个主题由一类句子来表示, 一个优秀的摘要应该要包含那些最重要的主题. 如今大部分的方法是建立一个模型来计算句子得分, 然后选择得分最高的部分句子来生成摘要. 不同于这些方法, 我们更加关注文本的主题而不是句子, 把如何生成摘要的问题看成一个主题的发现, 排序和表示的问题. 我们首次引入 dominant sets cluster(DSC)来发现主题, 然后建立一个模型来对主题的重要性进行评估, 最后兼顾代表性和无重复性来从各个主题中选择句子组成摘要. 我们在 DUC2005、2006、2007 三年的标准数据集上进行了实验, 最后的实验结果证明了该方法的有效性.

关键词: 多文本自动摘要; Dominant sets cluster

Query-Focused Multi-Document Summarization Based on Dominant Sets Cluster

LI Cheng-Guo

(Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
(University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Query-focused multi-document summarization aims at automatically creating a brief statement that presents the main points of a given document set and is relevant with the query. A given document set usually contains some themes. And each theme is represented by a cluster of sentences, and an excellent summary should cover the most important themes. Most of the existing multi-document summarization methods use a sentence-ranking model to select sentences to generate summary. These methods just consider cluster as a factor influences rank sentence or ignore it. Due to the influence of other factors, finally generated summary may not contain some important themes by these methods. Different from these methods, we focus on the themes level rather than sentence level and we treat the task as a themes detection, ranking and representation (TDRR) problem. We introduce dominant sets cluster (DSC) to produce theme clusters, construct a model to rank theme clusters, and select most representative and maximum information gain sentences to form summary. The experimental results on an open benchmark data sets from DUC05 to DUC07 show that our proposed approach is effectiveness.

Key words: multi-document summarization; dominant sets cluster; query-focused summarization

随着互联网的发展, 人们所缺乏的不在是信息, 而是如何有效的浏览和查阅信息的手段. 比如对于一个简单的关键词“文本摘要”[1-2], 谷歌搜索有约 700 万个相关的结果, 即使只浏览最相关的几十个结果, 对于用户来说也不是一个轻松的任务. 自动文本摘要技术对文本信息进行压缩表示, 可以帮助用户更好的浏览和吸收互联网上的大量信息. 由于技术方面的限制, 自动摘要和人工摘要之间还有比较大的差距, 至今也

①收稿时间:2014-03-18;收到修改稿时间:2014-04-14

没有成熟的商业系统，正因为如此，进一步激发了研究者是兴趣.

Tombros[4]研究了偏向查询的摘要在信息检索中的应用，得出了偏向查询的摘要比一般的搜索引擎提供的预览能够帮助用户更快的判断一个网页的相关性. 之后，偏向主题的文本自动摘要引发了更加广泛的关注，成为研究的一个热点.

1 相关工作

如今计算机自动生成摘要主要有两种方式，一种是抽取式摘要(extractive)，一种是抽象式摘要(abstractive). 前一种方法选择在原始的文本或者文本集中出现的句子组合成摘要；而后一种方式会利用一些自然语言处理的技术(如句子压缩、重构，指代消解等)，生成新的句子组成摘要. 尽管抽象式的摘要可能得到更加简洁明了的摘要，但是由于自然语言处理技术不够成熟，当前抽取式的摘要更加流行. 接下来我们介绍一些主流的抽象式摘要方法.

基于特征的方法：这类方法一般先定义一些特征，把每个句子用这些特征表示出来；然后建立模型或者用机器学习的方式来计算每个句子的得分；最后根据句子的得分来选择句子组成摘要. 一些经典的方法如MEAD[5]选择了句子位置、首句重叠度、centroid 值作为特征来计算句子重要程度. 其他一些研究表明词频、大小写、句子长度、TFIDF 等特征也能对有一定改善. 支持向量机(SVM)、隐马尔科夫(HMM)、条件随机场(CRF)等机器学习技术也被用于这一领域. 这类方法的优点是能够取得比较好的结果，缺点在于有的特征提取比较复杂，还有就是需要训练数据.

基于图的方法：这类方法一般先构建一个图，通常用图中的点来表示句子，图中的边则根据句子的相似程度来构造；然后用一些基于图的算法来计算句子的重要程度；最后选择最重要的那些句子组成摘要. 具有代表性的方法有 LexRank[6]、Manifold[1]等. 这类方法通常是无监督的，而且能取得较好的实验结果，正越来越受到关注.

基于矩阵分解的方式：这类方法一般从构建矩阵开始(如句子和词的矩阵、相似度矩阵)；然后利用矩阵分解的方法处理矩阵；最后根据处理矩阵得到的结果来选择句子生成摘要，比较有代表性的方法是 SVD[7].

其他方法：如基于冗余移除的 MMR[8](Maximal Marginal Relevance)、把自动摘要问题看成数据重构问题的 DSDR[9](document summarization based on data reconstruction)等.

总的来说，这些方法研究的重点都在句子，我们认为这样的方式可能导致最后生成的摘要不能很好的覆盖原来文本中的主题. 而我们的模型能够发现文本的主题，计算每个主题的重要程度，最后选择代表最重要的主题的句子组成摘要.

2 基于 DSC 的多文本自动摘要模型

2.1 概述

基于句子抽取的摘要方法选择原来文本中的句子来组成摘要，如何选择句子成为这一类方法的核心. GB6447-86 文摘编写规则对文摘的定义是：以提供文献内容梗概为目的，不加评论和补充解释，简明、确切地记述文献重要内容的短文[10]. 结合文摘的定义，我们选择的句子应该能够表达文本的主要内容，提供文本内容梗概；由于抽取式方法的特点，“不加评论和补充解释”的要求自动满足；同时文摘还要求简洁，也就是说代表同样内容的句子，我们不应该让它们重复的出现在摘要中. 综上，我们认为抽取的句子要能够表达文本的重要内容，同时尽量避免内容上的重复.

那么什么是文本的重要内容呢？一个给定的文本集通常包含一些主题，而且每个主题由一类句子来表达. 我们认为这些主题中最重要的那部分，就是文本的重要内容. 所以，我们把如何生成摘要的问题看成一个主题发现，排序和表示的问题. 如图 1 所示，我们的模型主要如下分成 3 个模块:

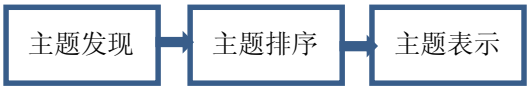

图 1 基于 DSC 的多文本自动摘要模型

1、主题发现：这个模块的目标是发现文本集中的主题. 我们用 DSC(dominant sets cluster)[3]来对句子进行聚类，然后认为每一类句子表示一个主题.

2、主题排序：这个模块的目标是对每个主题进行排序. 我们认为：一个主题中包含的句子越多越重要；DSC 算法先生成的主题比后面生成的主题更重要；和查询相关性越大的主题越重要. 据此我们建立了一个模型计算主题的得分，然后根据得分对每个主题进行

排序.

3、主题表示: 这个模块的目标是选择合适的句子来表示主题. 因为最后生成的摘要必须要避免内容上的重复, 这个模块我们从句子对一个主题的代表性和摘要的无重复性两个方面进行建模, 对句子进行选择.

2.2 主题发现

当前主要的主题侦测一般用聚类的方式来完成, 认为聚类后的每一个类就是一个主题, 常用的聚类方法有 k 均值聚类、基于矩阵分解的聚类(MF)等. 这些方法应用于这一领域的时候来说, 有个很大的缺陷: 必须先设置最后聚类的个数. 而对于一个文本集来说, 准确预测其中类的个数是很困难的. 为了克服这个缺陷, 我们引入了 DSC[3](dominant sets cluster)进行聚类, 完成主题发现.

我们先把文本集 D 拆分成一堆句子的集合 $D=\{s_1,\ldots,s_n\}$, 其中 s_i 代表一个句子. 我们定义句子的相似度矩阵为 $A=(a_{ij})$, 其中 $a_{ij}=\cos sim(s_i,s_j)$ 代表句子 i 和句子 j 的相似度. 这里的相似度用计算方式如下: 先将句子表示成向量, 向量的每一个值表示单词在句子中出现的次数; 然后计算句子向量的余弦相似度. 为了使用 DSC 算法, 对任意的 i, $a_{ii}=0$. DSC 算法每次循环得到一个点集 DS(dominant set)作为一个类, 然后将这些点移除, 重复寻找下一个 DS. 下面是具体的步骤:

1、设 $k=1$ 表示第一次循环.

2、设 $t=1$ 初始化向量, n 为剩余的句子个数.

3、按如下公式迭代直到收敛, 解得 X

$$x_i(t+1)=x_i\frac{(AX(t))_i}{X^T*AX(t)}$$

4、选择 X 中大于 0 的值对应的句子组成类 c_k, 选择 X 中大于 0 的值组成支持向量 v_k 表示这些句子属于 c_k 的概率.

5、在 A 中移除上一步中选择的句子对应的行和列, 更新循环次数 k 和剩余句子个数 n.

6、重复第二步到第五步直到剩余句子为 0.

完成以上步骤后, 我们得到 k 个类 $c_1,\ldots,c_k$ 表示 k 个主题, 以及对应 k 个向量 $v_1,\ldots,v_k$ 表示这些主题中的句子属于该主题的概率.

2.3 主题排序

我们认为一个类代表文本集中的一个主题, 但是一个文本的主题有主次之分, 这部分我们将讨论如何判断一个主题的重要程度.

很显然, 一个主题如果被重复提到多次, 那么这个主题肯定比提到次数少的主题重要. 但是我们并不能简单的用一个主题中的句子个数来表示这一重要程度, 因为有可能一个主题中的某些句子并没有表达这一主题. 我们认为一个主题中句子的相似程度可以在一定程度上代表这种影响, 相似程度越高, 表明这些句子代表这个主题的程度越高, 反之亦然. 通过以上分析, 我们根据一个类的内部信息建模, 得到一个得分表示这个主题在这方面的重要程度, 具体计算方式如下:

$$S_{\text{int er}}=avgsim(c_i)*\log 2(sizes(c_i))$$

其中 $avgsim(c_i)$ 是一个主题中所有句子的平均相似程度, $sizes(c_i)$ 是一个主题中包含的句子个数.

DSC 算法每次循环生成 DS(dominant set)作为一个类, 前面生成的类比后面生成的具有更好的性质, 即内部的高相似性和外部的低相似性. 我们根据生成主题的顺序, 用如下计算方式表示这一得分:

$$S_{order}=\alpha^{(1/i)}$$

其中 α 是参数, 而 i 表示表示该主题是第 i 个生成的.

对于偏向查询的摘要, 在计算主题重要程度的时候, 我们除了考虑一个主题本身的性质, 还应该考虑这个主题和查询的相关程度. 这里我们把一个主题中平均每个句子中包含的查询词个数作为这方面的得分, 具体如下:

$$S_{query}=querrwords(c_i)/sizes(c_i)$$

其中 $querywords(c_i)$ 是一个主题中所有句子包含的查询词个数, 而 $querywords(c_i)$ 是一个主题中包含的句子个数.

最后我们把综合考虑以上三个因素, 把三个得分加起来作为一个主题最后的得分:

$$S_{c_i}=S_{\text{int er}}+S_{order}+S_{query}$$

然后我们根据每个主题的得分进行排序, 同时调整支持向量使之对应.

2.4 主题表示

我们假设根据上一模块排序后的主题是 $c_1,\ldots,c_k$, 其中 c_1 是得分最高的主题, 以及对应 k 个向量 $v_1,\ldots,v_k$. 一个很直观的方式是先从 c_1 中选择最具有代表性的句子加入摘要, 然后再从 c_2 中选择最具有代表性的句子加入摘要, 这样进行下去直到摘要的长度达到指定值. 但是这样选择的句子可能会有些信息

上的重复，一个好的摘要系统应该避免这一点. 所以我们综合考虑句子的代表性和无重复性进行选择.

我们首先考虑哪个句子更能代表一个主题. 在 DSC 聚类完成之后，对于每一个主题，我们得到一个支持向量，该向量的每一个值可以理解成对应的句子属于这个主题的概率. 一个句子属于某个主题的概率越高，显然这个句子越能代表这个主题，所以这我们用这个概率来表示一个句子对于该主题的代表性. 为了满足无重复性的要求，我们借鉴了[11]中提出的方法，在每次选择完一个句子之后，对这个句子包含的词的权重进行惩罚. 具体过程如下:

1、设 $t = 1$，计算每一个词 w_i 在文本集中的概率分布 $p_t(w_i) = n / N$，其中 n 表示词 w_i 在文本集中出现的次数，N 表示文本集的总词数.

2、对 c_t 中的每一个句子 s 根据如下公式计算得分:

$$Score(s) = \varphi * v_t(s) + (1 - \varphi) * novelty(s)$$

其中，$novelty(s)$ 是句子 s 中所有词的平均概率，φ 是值为 0 到 1 的参数，用来调节代表性和无重复性的比重.

3、选择得分最高的句子 s 加入摘要，同时对句子 s 中出现的所有词的权重按如下方式进行惩罚:

$$p_{t+1}(w_i) = p_t(w_i) * p_t(w_i)$$

4、如果摘要达到指定长度就停止，否则 $t = t + 1$，返回第二步继续选择句子.

3 实验结果和分析

3.1 数据集

我们采用 DUC(Document Understanding Conference)提供的测试数据来进行试验. DUC 是文本自动摘要领域权威的国际会议，其提供的测试数据被广泛使用，DUC2005、2006、2007 都提供了基于查询的文本自动摘要测试数据集. 其中每个数据集包含 45-50 个查询，每个查询对应 25-50 篇文本，这些文摘都来至于一些国际权威的新闻通稿，并且都和查询相关，对每个查询还包含四个由专家给出的参考摘要，可以作为评价自动文本摘要好坏的标准. 对每个查询希望能够生成长度不超过 250 个词的摘要. 下表给出这三个数据集的简要信息.

表 1 DUC 数据集

	DUC2005	DUC2006	DUC2007
查询个数	50	50	45
文本个数	25-50	25	25
文章来源	金融时报，洛杉矶时报	联合通讯社，纽约时报，新华社	联合通讯社，纽约时报，新华社

3.2 评价标准

DUC 采用 BE、ROUGE、Pyramid 三种自动评价方式对自动生成的摘要进行评价，目前 ROUGE 是国际上普遍采用的评价方式，我们采用 ROUGE 1.5.5[12] 作为自动评价的标准. ROUGE 将 n 元词作为基本单元，把人工摘要作为标准，计算自动摘要中 n 元词的召回率、准确率及综合二者的 F 值. 在我们的实验中，采用和 DUC 一样的 ROUGE 参数设置: “-n 4 –w 1.2 –m -2 4 –u –c 95 –r 1000 –f A –p 0.5 –t 0 –a –d”.

ROUGE-n 表示以 n 元词作为基本单元的评价方式，ROUGE-1 被这一评价标准被认为和人的评价相似度最高，在我们的实验中选择 ROUGE-1 和 ROUGE-2 的 F 值作为评价指标.

3.3 实验

我们实现了以下方法和我们的方法进行比较:

LSA[7]: 用 SVD 分解得到文本的隐含主题，和每个句子在主题中的得分，依次从每个主题中选择得分最高的句子组成摘要.

SumBasic[11]: 先计算单词在文本集中的概率，然后用句子中所有词的概率之和作为句子得分，依次选择得分最高的句子组成摘要，同时用每次选择句子后降低句子中包含的词的权重的方式来保证摘要的无重复性.

DSDR[9]: 将自动摘要问题看成用摘要来重构文本集的重构，选择使得重构误差最小的句子集合作为摘要.

Manifold[1]: 用流形排序的方式来计算句子和查询的相关程度，以此作为句子的得分，选择高分句子组成摘要.

LexRank[6]: 用点表示句子，相似度在一定阈值上的点用边连接起来，再用马尔科夫随机游走模型来计算每个点的得分，最后选择高分句子组成摘要.

NCBSum[13]: 考虑生成的摘要应该满足重要性，无重复性，全面性，平衡性，根据这四点进行建模，计

算句子得分, 最后选择高分句子组成摘要.

对算法中的参数, 我们根据 DUC2007 的数据进行了经验性的设置, 具体 的: $\alpha = 20, \varphi = 0.8$.

表 2、3、4 给出了实验结果, 可以看出我们的方法在所有数据集上的两个评价标准上都优于其他方法, 证明了从主题层面上来考虑如何生成摘要, 确实能够得到比较好的效果.

表 2 DUC2007 上的实验结果

Method	Rouge-1	Rouge-2
LSA	0.3855	0.0816
SumBasic	0.3921	0.0740
DSDR	0.4159	0.1026
Manifold	0.4053	0.1013
LexRank	0.4057	0.1048
NCBSum	0.4127	0.1041
Ours	**0.4294**	**0.1079**

表 3 DUC2007 上的实验结果

Method	Rouge-1	Rouge-2
LSA	0.3647	0.0673
SumBasic	0.3769	0.0610
DSDR	0.3909	0.0832
Manifold	0.3892	0.0825
LexRank	0.3613	0.0686
NCBSum	0.3887	0.0828
Ours	**0.4035**	**0.0884**

表 4 DUC2007 上的实验结果

Method	Rouge-1	Rouge-2
LSA	0.3205	0.0488
SumBasic	0.3400	0.0489
DSDR	0.3371	0.0613
Manifold	0.3690	0.0751
LexRank	0.3367	0.0549
NCBSum	0.3725	0.0725
Ours	**0.3745**	**0.0746**

4 结语

互联网的发展给我们提供了海量信息, 但是如何有效的利用这些信息成为一个需要解决的问题. 自动文本摘要技术可以将大量文本压缩成指定长度的短文, 同时尽量保留其中的重要信息, 可以有效解决信息过载的问题. 如今的主流研究方式是从句子层面上进行分析, 如何对句子进行打分及排序, 而我们从文本主题这一层进行分析, 建模, 最后通过实验证明了我们方法的有效性. 我们在下一步准备研究聚类得到的主题和人工得到的主题的关系, 以及考虑各主题之间的关系和对最后生成的摘要的影响. 我们还准备使用自然语言处理技术来生成抽象式的摘要.

参考文献

1 Wan X, Yang J, Xiao J. Manifold-ranking based topic-focused multi-document summarization. Proc. of IJCAI. 2007,7. 2903–2908.

2 Ouyang Y, Li W, Li S, Lu Q. Applying regression models to query-focused multi-document summarization. Information Processing & Management, 2011, 47(2): 227–237.

3 Pavan M, Pelillo M. Dominant sets and pairwise clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007, 29(1): 167–172.

4 Tombros A, Sanderson M. Advantages of query biased summaries in information retrieval. Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998. 2–10.

5 Radev D, Jing H, Stys M, Tam D. Centroid-Based summarization of multiple documents. Information Processing & Management, 2004, 40(6): 919–938.

6 Erkan G, Radev DR. Lexrank: Graph-based lexical centrality as salience in text summarization. Artif. Intell. Res. (JAIR), 2004, 22. 457–479.

7 Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis. Proc. of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2001, 19–25.

8 Carbonell J, Goldstein J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998. 335–336.

9 He Z, Chen C, Bu J, Wang C, Zhang L, Cai D, He X. Document summarization based on data reconstruction. 26th AAAI Conference on Artificial Intelligence. 2012.

10 文摘编写规则, gb6447-86.

11 Nenkova A, Vanderwende L. The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101, 2005.

12 Lin CY. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proc. of the ACL-04 Workshop. 2004. 74–81.

13 Li L, Zhou K, Xue GR, Zha H, Yu Y. Enhancing diversity, coverage and balance for summarization through structure learning. Proc. of the 18th international conference on World wide web. ACM, 2009. 71–80.