基于交互增强原理的多文档自动文摘算法*

王小磊[1,2]，张瑾[1,2]，许洪波[1]

1. 中国科学院 计算技术研究所 信息智能与信息安全研究中心，北京 100190

2. 中国科学院 研究生院，北京 100039

E-mail: wangxiaolei@software.ict.ac.cn, zhangjin@software.ict.ac.cn, hbxu@software.ict.ac.cn

摘要：本文提出了一种基于交互增强原理的多文档自动文摘方法。首先对句子集合和文档集合建立二部图，然后根据交互增强原理计算每个句子和文档的重要性得分。为了去除冗余，用 Normalized-Cut 方法将句子聚类成几个不同的子主题，并选出重要性得分最高且不在同一子主题中的句子生成文摘。最后，在 DUC2007 测试数据上通过实验证明了本文所提出方法的有效性。

关键词：二部图，交互增强原理，重要性得分，Normalized-Cut

Multi-document Summarization using Mutual Reinforcement Principle

Xiao Lei[1,2], Jin Zhang[1,2], Hongbo Xu[1]

1. Research Center of Information Intelligence and Information Security, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190

2. Graduate University of Chinese Academy of Science, Beijing 100039

E-mail: wangxiaolei@software.ict.ac.cn, zhangjin@software.ict.ac.cn, hbxu@software.ict.ac.cn

Abstract: In this paper, we propose a novel Multi-document Summarization method using Mutual Reinforcement Principle. Firstly, we build an bipartite graph for the set of sentences and the set of documents. Then we compute salience scores of each sentence and of each document using Mutual Reinforcement Principle. Furthermore, we cluster the sentences into different sub-topics with Normalized-cut method, and we extract those sentences which have highest salience scores and are in different sub-topics to generate summaries. The results evaluated on dataset of DUC 2007 demonstrate our proposed approach can perform well.

Key words: bipartite, mutual reinforcement principle, salience score, normalized-cut

1 引言

自动文摘技术用于自动从一篇或多篇文章中提取满足用户或应用需求的内容，加以组织后生成一篇内容完整、形式严谨的自动文摘。单文档文摘技术可以帮助人们在海量信息中准确、高效地寻找自己需要的信息，发展至今，已经得到了广泛的应用。但随着互联网的普及，现有的这些方法已经不能满足人们新的需求，在对多篇同一主题文档进行汇总和压缩的问题上仍然存在一些方法上的不足。多文档文摘技术则应运而生，它是信息时代发展到一定程度的必然趋势。多文档

* 本文承国家 973 “大规模文本内容计算”课题（2004CB318109）的资助。

文摘可以将多篇同一主题的文档进行汇总，让其中多次重复的信息一次出现在文摘中。多文档文摘的研究为用户提供了方便，提高了用户获取信息的速度和效率，为互联网的应用开辟了新的方向。

真正通用领域的多文档文摘研究是从 1997 开始的。近些年来，国际上一些比较权威的 NLP 领域会议，例如 ACL[1]、COLING[2]和 SIGIR[3] 等都有自动文摘的专题，另外由 NIST 支持的 DUC 会议 (Document Understanding Conference)[4] 是专门进行自动文摘技术的讨论与研究的，已经连续进行了很多年，使研究者共同参与到大规模文本测试中来，促进了自动文摘技术的发展。

本文提出一种基于交互增强原理(Mutual Reinforcement Principle，MRP)的多文档自动文摘方法。因为句子的重要性和文档的重要性是相互关联的，互为因果，构成一个循环，所以可以用迭代方法计算句子的重要性得分，并根据得分进行文摘句抽取。本文的方法分三个阶段：1) 根据交互增强原理计算句子的重要性得分；2) 利用 Normalized-Cut[6]将句子划分成不同的子主题；3) 选取重要性得分最高且分属于不同子主题的句子生成文摘。

2 相关工作

目前，大多数研究都致力于句子抽取型的文摘方法，即直接选取句子作为文摘句。利用统计学方法和启发式规则，对所有句子进行重要度打分，作为对主题的反应程度，然后选取重要性得分最高的句子生成文摘。Salton 首先提出了中心度 (Centrality) 的概念[1]，利用句子的中心度得分作为选取文摘句的依据。Moens 等人[2]用余弦相似度将句子聚类到不同的主题区域，然后选择与同一聚类中的句子的相似度之和最大的句子作为文摘句。Zha 对词和句子建立二部图[3]，如果某个词出现在一个句子里，则二者之间有一条边。 他运用了交互增强原理 (Mutual Reinforcement Principal)： 一个词如果出现在很多重要性得分很高的句子中，则其重要性得分也很高；反之，如果一个句子包含很多重要性得分很高的词，则其重要性得分亦很高。求解重要性得分最终可以转化为求取二部图对应的转移矩阵的奇异向量问题。Erkan 和 Radev 提出了 LexRank 方法[4]。他们将 PageRank[5]算法的思想引入到自动文摘中。在他们定义的图中，每个点代表一个句子，如果两个句子的相似度大于某一临界值，则有一条边连接这两个句子。每个句子的重要性由与它相邻的所有句子的重要性决定。即一个句子如果与很多重要度很高的句子相邻，则它本身的重要度也很高。

本文提出一种新的基于交互增强原理的多文档自动文摘方法，利用交互增强原理对句子按照重要性排序。因为词的重要性与语义环境紧密相关，变化很大。所以 Zha 通过词和句子间相互关系，利用交互增强原理计算出的重要性得分意义不明确。而句子和文档的含义和重要性相对稳定，可以通过重要性得分来反映。因此本文首先对句子和文档建立二部图，利用交互增强原理，求取句子的重要性得分。其中，两种不同形式的交互增强原理不分别计算句子的重要性得分，并对结果进行了比较。为了去除冗余，本文用 Normalized-Cut 方法[6]将句子聚类成不同的子主题，最后选出重要性得分最高且不在同一子主题中的句子生成文摘。

---

[1] ACL: http://www.aclweb.org/
[2] COLING: http://www.dcs.shef.ac.uk/research/ ilash/iccl/
[3] SIGIR : http://www.sigir.org/
[4] DUC : http://duc.nist.gov/

3 基于交互增强原理的自动文摘算法

3.1 句子与文档相似度计算

为了对句子和文档建立二部图，首先要计算句子和文档间的相似度。本文采用传统的向量空间模型，将句子表示为词空间中的向量。首先产生两个集合：所有句子集合 $S=\{s_1, \cdots, s_n\}$ 和所有文档集合 $D=\{d_1, \cdots, d_n\}$。如某个句子 $s_j=\{t_{1j}, \cdots, t_{lj}\}$，其中 l 为特征词总数，t_{ij} 为某个词 t_i 在句子 s_j 中的权重，加权方式为 tf-isf，具体定义如下：

$$t_{ij} = tf_{ij} * \log \frac{n}{sf_i} \tag{1}$$

其中，tf_{ij} 是 t_i 在句子 s_j 中出现的频数；sf_i 是整个文档集中包含 t_i 的句子数；n 是整个文档集中的句子总数。

对于文档，实际上它不仅是一个长句子，而是由句子构成的集合。所以某个句子 s_i 和某个文档 d_j 间相似度定义为 s_i 与 d_j 中所有句子的相似度之和，表示为：

$$w_{ij} = \sum_{s_k \in d_j} sim(s_i, s_k) \tag{2}$$

3.2 基于交互增强原理的自动文摘算法

如果某个句子 s_i 和某个文档 d_j 间相似度 w_{ij} 不为 0， 则用一条边连接 s_i 和 d_j，边的权重为 w_{ij}。文档-句子二部图可以表示为 $G=\{S, D, W\}$。其中 S 和 D 分别为句子集合和文档集合。$W=(w_{ij}>0)$ $(i=1, \cdots, n;\ j=1, \cdots, m)$ 为句子-文档相似度矩阵。

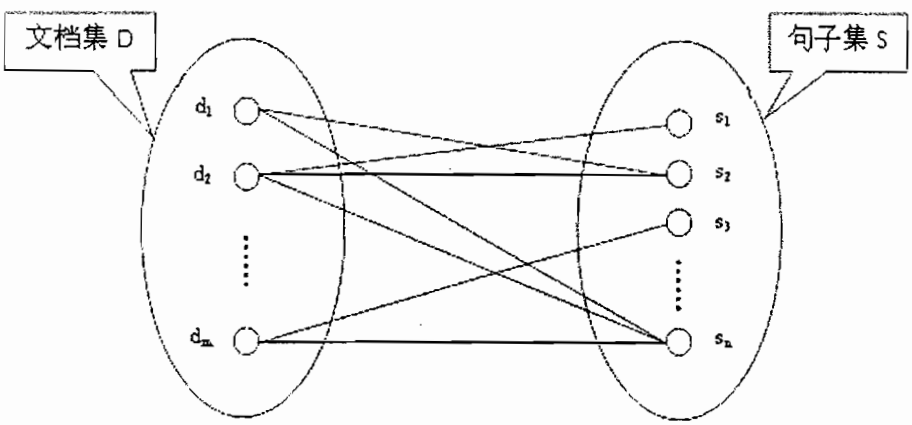

表 1 文档-句子二部图

句子连接度矩阵可以定义为：

$$P = diag\{p_1, \ldots, p_n\} \tag{3}$$

其中 $p_i = \sum_{j=1}^{m} w_{ij}$ 为句子 s_i 与所有文档的相似度之和。而文档连接度矩阵也可以类似定义为：

$$Q = diag\{q_1, \ldots, q_m\} \tag{4}$$

其中 $q_j=\sum_{i=1}^{n} w_{ij}$ 为文档 d_j 与所有句子的相似度之和。

本文利用两种不同形式的交互增强原理计算每个句子 s_i 和每个文档 d_j 的重要性得分。第一种表述为：

如果与一个句子相关的所有文档的重要性得分很高，则这个句子的重要性得分也很高。反之，如果与一个文档相关的所有句子的重要性得分很高，则这个文档的重要性得分也很高。

上述原理可以形式化表示为：

$$u_i \propto \sum_{j=1}^{m} w_{ij} v_j \tag{5}$$

$$v_j \propto \sum_{i=1}^{n} w_{ij} u_i \tag{6}$$

其中 u_i 为句子 s_i 的重要性得分；v_j 为文档 d_j 的重要性得分。从等式（5）中可以看出，句子 s_i 的重要性得分为：所有文档的重要性得分乘以它们与 s_i 的相似度，最后求和。文档的重要性得分类似。通过矩阵形式可表示为：

$$u=\frac{1}{c_u}Wv \tag{7}$$

$$v=\frac{1}{c_v}W^T u \tag{8}$$

其中 $u=(u_1,\cdots,u_n)^T$ 为句子重要性得分向量；$v=(v_1,\cdots,v_n)^T$ 为文档重要性得分向量。c_u 和 c_v 分别是 u 和 v 的标准化因子，满足 $\|u\|=\|v\|=1$。具体求解算法为 MRP1：

1) 将 v 初始化为 $(1,\cdots,1)^T$；
2) 根据 $u=Wv$ 计算得到 u，并标准化：$u=u/\|u\|$；
3) 根据 $v=W^T u$ 计算得到 v，并标准化：$v=v/\|v\|$；
4) 重复 2)和 3）直到收敛。

第二种形式的交互增强原理，其形式化表示为：

$$u_i=\sum_{j=1}^{m}\frac{w_{ij}}{q_j}v_j \tag{9}$$

$$v_j=\sum_{i=1}^{n}\frac{w_{ij}}{p_i}u_i \tag{10}$$

从等式(9)中可以看出，文档 d_j 的重要性得分 v_j 按照一定比例 w_{ij}/q_j 分配给所有与它相关的句子。显然 w_{ij} 越大，则句子 s_i 从文档 d_j 分配到的重要性得分也越多。所有文档分配给句子 s_i 的重要性得分之和即为 s_i 的重要性得分 u_i。文档的重要性得分类似。通过矩阵形式可表示为：

$$u=WQ^{-1}v \tag{11}$$

$$v=W^T P^{-1}u \tag{12}$$

具体求解算法为 MRP2：

1) 将 v 初始化为 $(1/m,\cdots,1/m)^T$；

2) 根据 $u=WQ^{-1}v$ 计算得到 u；

3) 根据 $v=W^TP^{-1}u$ 计算得到 v；

4) 重复 2)和 3) 直到收敛。

在文摘长度限制范围内，重要性得分高的句子入选文摘句。本文对这两种使用不同形式的交互增强原理的算法 MPR1 和 MPR2 进行了对比。实验证明，后者的效果更好。

3.3 利用 Normalized-Cut 划分子主题

直接抽取重要性得分最高的句子生成的文摘内包含大量冗余信息。为了使重复的信息在文摘中只出现一次，本文使用 Normalized-Cut 方法将句子聚类成不同的子主题，每个子主题中的内容近似相同。在生成文摘时，只从每个子主题中选取一个句子。Normalized-Cut 是一种基于图的聚类方法。假设句子间相似度矩阵为：$M=(m_{ij})(i,j=1,\cdots,n)$，其中 $m_{ij}=sim(s_i,s_j)$ 为句子 s_i 和 s_j 间相似度。相应的图表示为 $G=\{S,M\}$。如果句子 s_i 和 s_j 间相似度不为 0，则 s_i 和 s_j 间有边连接，权重为 m_{ij}。句子连接度矩阵可以定义为：

$$R = diag\{r_1,\ldots,r_n\} \tag{13}$$

其中 $r_i=\sum_{j=1}^{n} m_{ij}$ 为句子 s_i 与所有句子的相似度之和。通过对图进行分割即聚类，可以得到多个子图。一种简单的想法是找到一种图的分隔方法使连接不同子图内结点的边的权重之和最小。而这样得到的图分割结果往往不令人满意。可以想象，如果很多子图仅包含一个结点，则子图间的边的权重之和一定很小。解决这个问题的一种方法是限制各个子图中的结点数足够多。由这种思想产生了两种算法：Ratio-Cut [7]和 Normalized-Cut[6]。Ratio-Cut 考虑子图中结点数目的平衡。而 Normalized-Cut 不仅考虑各聚类内的结点数目的平衡，还考虑了结点的重要性的平衡。文献[8]证明了 Normalized-Cut 方法可以转化为下面的特征值求解问题。即求解

$$(R-M)x = \lambda Rx \tag{14}$$

的特征值，按由小到大排序为 $\lambda_1,\cdots,\lambda_n$，其中 $\lambda_1=0$。然后取 $\lambda_2,\cdots,\lambda_{k+1}$ 对应的特征向量 $x_2,\cdots,x_{k+1}$，并由此得到每个句子的向量表示：$s_i=\{x_{2,j},\cdots,x_{k+1,j}\}$。最后可以使用 k-means 对这些向量进行聚类，便可得到最优解。

通过对句子聚类将它们划分为不同的子主题，最后抽取重要性得分最高且不在同一子主题中的句子，生成文摘。

4 实验结果及分析

自 2001 年以来，DUC 每年都通过一系列的文摘任务来对比不同研究组的文摘系统的优劣。DUC 2007 主任务提供了 45 个文档集，每个文档集包含 25 篇文章和一个简短的主题说明，要求生成与主题相关的文摘。此外，每个文档集都提供了 4 篇标准文摘用于评测。本文利用 DUC 2007 主任务的测试语料集进行实验，并利用文摘评测领域著名的 ROUGE 工具[9]进行评价。ROUGE 首先由多个专家分别生成人工文摘，构成标准文摘集。然后将提交的待评测文摘与标准文摘作对比，通过统计二者之间匹配的基本单元（n 元语法、词序列和词对）的个数来评价文摘质量。通过多

专家人工文摘的对比，提高评价系统的稳定性和健壮性。ROUGE 主要包括以下四种评价标准：1）ROUGE-N 基于 n-gram 共现统计；2）ROUGE-L 基于最长公共子串统计；3）ROUGE-S 基于顺序词对统计；4）ROUGE-W 在 ROUGE-L 的基础上，考虑串的连续匹配。在此，本文选用 ROUGE-2 和 ROUGE-SU4 这两种主流指标进行评价。

为了评估算法的有效性，我们将自己的方法与 Zha 的方法以及 DUC 2007 的两个 baseline 做对比。其中 Zha 的方法是对词和句子建立二部图，并根据第一种形式的交互增强原理求解句子的重要性得分，然后选取得分高的句子作为文摘句。DUC 2007 的 Simple baseline 通过选取每篇文章的段首句生成文摘。另一个 Generic baseline 直接使用稍作修改后 CLASSY（参与 DUC 评测的某个系统）的评测结果。我们在对文档和句子建立二部图后，利用了两种不同形式的交互增强原理 MPR1 和 MPR2 分别计算句子的重要性得分。在文摘长度限制范围内（不超过 250 个单词），选择得分最高的句子生成文摘。评测结果如表 1 所示。

评测结果	ROUGE-2	ROUGE-SU4
Simple baseline	0.06039	0.10507
Zha	0.07529	0.13280
Generic baseline	0.09382	0.14641
MRP1	0.09600	0.14903
MRP2	**0.10574**	**0.15966**

表 1 ROUGE 评测结果比较

从表 1 可以看出，MRP1 和 MRP2 的 ROUGE-2 和 ROUGE-SU4 的评测结果明显高于 Zha 的算法和两个 baseline。其中，MRP2 的结果最好。

为了去除冗余，在 MRP2 算法的基础上，本文利用 Normalized-Cut 对句子聚类成不同子主题。聚类数目设定为待生成文摘的文档集中的文档数目（25），聚类中心初始值设定为重要性得分最高的 25 个句子。最后选取重要性得分最高且不在同一子主题中的句子作为文摘句。考虑到有些聚类中的句子相对于整个文档集不具有代表性，所以我们又加入了新的限制，即仅选择重要性得分排名在 K(=20, 30, 40) 位之前的句子，在重要性得分和聚类间优先考虑前者。具体评测结果如表 3 所示。

评测结果	ROUGE-2	ROUGE-SU4
MRP2	0.10574	0.15966
K=20	**0.10962**	**0.16373**
K=30	0.10933	0.16246
K=40	0.10963	0.16281

表 2 引入聚类后评测结果比较

进行聚类后的评测结果有一定提高。当 K=20 时结果最好：ROUGE-2 得分 0.10962，ROUGE-SU4 得分 0.16373。

5 结论

本文提出了一种新的基于交互增强原理的多文档自动文摘算法。首先，我们根据文档和句子间的相似度对它们建立二部图，得到它们之间的一种关系表示。句子的重要性得分根据交互增强原理计算得出，可以反映句子在文档集中的重要程度。其中，我们使用了两种不同形式的交互增强原理来计算句子的重要性得分，在 DUC 2007 上的评测结果表明这两种算法都具有较好的性能。为了进一步改进结果，本文还引入了 Normalized-Cut 去除冗余，最终生成较好的文摘。以后我们将对算法作进一步改进：1) 在计算句子与文档相似度时引入更多信息，比如句子在文档中的位置以及句子的类型等；2) 引入聚类技术后，总体评测结果提升不很明显。我们将引入动态聚类思想，而不固定聚类数目；3) 改进最后的文摘句选取标准。希望通过上述改进能得到更好的效果。

参考文献

[1] G. Salton, A. Singhal, M. Mitra, & C. Buckley. Automatic Text Structuring and Summarization. Information Processing & Management, 33 (2), 1997, pp.193-207.

[2] M.-F. Moens, C. Uyttendaele, & J. Dumortier. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. Journal of the American Society for Information Science, 50(2), 1999, pp.151-161

[3] H. Zha. Generic Summarization and Key Phrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Tampere, Finland, 2002.

[4] Güneş Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. Journal of Artificial Intelligence Research (JAIR), 2004.

[5] L. Page, S. Brin, R. Motwani, & T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford, CA, 1998.

[6] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence, 22(8), 2000, pp.888–905.

[7] L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on CAD, 11, 1992, pp.1074–1085.

[8] Ulrike Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17, 2007, pp.395 – 416.

[9] C. Y. Lin. ROUGE: a package for automatic evaluation of summaries. In Proceedings of the ACL 2004 Workshop on Text Summarization, Spain, 2004.7, pp.4-8.