第38卷 第16期
Vol.38 No.16

计 算 机 工 程
Computer Engineering

2012年8月
August 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)16—0170—04 文献标识码：A 中图分类号：TP301.6

基于指代消解和篇章结构分析的自动摘录算法

郑 诚，刘福君，李 清

(安徽大学计算机科学与技术学院，合肥 230039)

摘 要： 传统自动文摘方法生成的文摘结果指代关系模糊，且对于某些段落结构有规律的文章，没有分析文章结构与主题思想之间的关系。为此，提出一种基于指代消解和篇章结构分析的自动摘录算法。采用有限知识的思路完成指代消解，利用指代消解解决文摘语义不连贯问题，以提高句子权重计算的准确性，对文章做主题划分时进行篇章结构识别，按照段落标题信息划分段落结构。实验结果表明，该算法在受限金融领域文本自动摘录中，具有较高的准确率和召回率。

关键词： 自然语言处理；自动摘录；向量空间模型；主题划分；篇章结构；指代消解

Automatic Extraction Algorithm Based on Anaphora Resolution and Text Structure Analysis

ZHENG Cheng, LIU Fu-jun, LI Qing

(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

【Abstract】 There are some problems should be considered in automatic extraction of traditional methods: Conference relations in the result of automatic extraction are not clear, some relationships between obvious structures of paragraphs and the theme of the text are not paid enough attention. For which, this paper presents a method based on anaphora resolution and text structure analysis, which combines the traditional statistics with regulars on automatic abstract. This method applies limited knowledge to pronoun resolution, which is to solve the problem of semantic incoherence, also to improve the precision when computing sentences' weight. Based on sequential paragraphic similarity, this method can recognize obvious topics to partition text. Experimental results show that this method improves precision and recall when it is applied for limited-financial field.

【Key words】 Natural Language Processing(NLP); automatic extraction; Vector Space Model(VSM); topic segmentation; text structure; anaphora resolution

DOI: 10.3969/j.issn.1000-3428.2012.16.044

1 概述

随着自然语言处理(Natural Language Processing, NLP)技术的发展，分词与词性标注技术已发展到相当成熟的地步，目前大部分词性标注工具对文本的标注正确率都达到了96%以上，解决了汉语句子分析中的层次问题，将含有一定意义的句子划分成具有最基本语义的元素，有利于将句法分析的复杂度各个击破，从而促进了自然语言处理的发展，如基于组块的研究、指代消解的研究方法。

文献[1]将现有各种自动文摘技术方向归结为自动摘录、基于理解的自动文摘、信息抽取和基于结构的文摘。基于上述观点，结合中文文本特点，中文自动文摘研究方法分类如下：(1)基于语料库方法，是一种改进的具有一定语义信息处理的字频统计方法。(2)用概念模型进行信息抽取的方法，主要采用 Ontology 描述应用领域的元素，并形成领域概念树来描述领域元素之间的关系。(3)基于Ontology 的智能信息提取方法。(4)利用知识库提取文本信息的方法。(5)结合文本语义的形式化模型，即建立语境框架。(6)分析篇章多级依存结构提取中心成分，采用该方法已研发出 HIT-863Ⅱ系统。

自动摘录的方法综合利用词频、标题、位置、句法结构、线索词和指示性短语等特征[2]的有机结合，考虑到文本形式的规律，能够适用于非受限域。但是由于各种文章的特征不一定符合同样的规律，因此该方法生成的摘要存在着反映主题不全面、主题冗余，以及语义不连贯的问题。

本文提出一种基于指代消解和篇章结构分析的自动摘录算法。针对金融领域，采用基于统计和规则的自动摘录方法，通过基本的指代消解方法，解决自动摘录方法面临的问题[3]。

2 指代消解处理

由于文摘候选句是根据句子权重将句子从文中的不

基金项目： 安徽省自然科学基金资助项目(11040606M133)

作者简介： 郑 诚(1966－)，男，副教授、博士，主研方向：语义信息检索，数据挖掘；刘福君、李 清，硕士研究生

收稿日期： 2011-10-20 **修回日期：** 2011-12-05 **E-mail：** liufujun860324@163.com

同位置抽取出来，在语义上具有一定的跳跃性，造成意义上和语句形式上的不连贯。在文本预处理时就进行指代消解的处理，沿用文献[4]“有限知识”的代词消解的方法。先行词选取指示符的打分范围为[−1,0,1,2]，包括以下方面：

(1)明确性：回指词[5]前面句子中的名词词组有“这”、“这些”等和所有格的，打分为 0；否则打分为−1。

(2)指示词：存在于动词词性的指示词词表{预计、预测、分析等}中的词语，其后第 1 个名词或名词词组是首选先行词(打分为 1)。

(3)词语重复：在一个段落里，如果名词词组重复 2 次以上，则打分为 2；重复一次，打分为 1；否则打分为 0。

(4)标题：如果一个名词词组出现在文章标题中，则打分为 2；若出现在段落标题中，则打分为 1；否则打分为 0。

(5)无介词前置的名词词组(该方法与中心理论一致)：名词词组前有介词的，打分为−1；对于没有介词的，打分为 0。

(6)搭配模式选择：同一搭配模式定义为(名词词组/代词，动词)、(动词，名词词组/代词)2 种模式。

(7)最近距离引用：启发式规则为：“……(代词)+动词 V_1+名词词组……+连词+(代词)+动词 V_2+它(+连词+(代词)+动词 V_3+它”，这里的连词是{及/并/并且/之后/之前/然后……}中的词。这里，V_1 后面紧跟着的名词词组有可能就是紧跟在 V_2 后的代词 it 的先行词。满足以上规则，则名词词组打分为 2；否则打分为 0。

(8)指称距离：选出回指词前面的 3 个候选先行词，分别根据其间距离打分为[−1,0,1]。距离以回指词与候选先行词之间的字数来衡量。

根据金融领域文章的特点，文中通常描述的是某个证券或政策的情况或影响，重点处理“这”、“此”、“这些”、“这类”等类似的回指词。有时，“这”和“此”代表的并不是某个词语，而是前面的一个观点、事件、或者情况等。具体算法如下：

(1)扫描当前句子与其前面 2 个句子(如果该句子段落中的不是首句或第 2 句话)，判断出回指词左边的名词短语。

(2)如果回指词是“这”，且词性标注为“PN”则将其标注为为无指代关系的限定词 Determiner。

(3)将步骤(1)找到的名词短语与相应的回指词进行性别与数目的一致性比较，如果一致，则将这些名词短语视为候选先行词。

(4)对每一个候选先行词进行先行词选取指标的使用并打分，得分最高的候选先行词作为选定的先行词。若 2 个候选先行词的分数一样高，则最近距离引用分数高的那个候选先行词选定为先行词；若它们都没有计算指称距离这个指标，则根据同一搭配模式分数高的作为先行词；如果同一搭配模式分数相等或者也没有计算到该指标，那么选取指示词指标高的作为先行词；若还是无法判定，则选取与回指词最近的作为先行词。

(5)标记为 Determiner，并且字符为“这”的不做处理，其他的回指词均替换为相应的先行词。

3 特征组合在金融领域文摘系统中的应用

特征组合在金融领域文摘系统中的应用如下：(1)标题特征，含有标题词语的句子通常可以表达文章的一部分中心思想。(2)位置特征，首段、尾段、段落首句和尾句含主题句很可能就是主题句。(3)线索词特征：根据金融领域的文章，总结一些线索词如预测、分析等。(4)指示词特征：文献[1]指出，可以利用指示性词语作为文摘句的选取指标。指示性词语可以提示后面的句子是对文中主题思想进行总结，如“综上所述”、“笔者认为”，含有这些短语的句子将被直接选为文摘句。(5)金融词汇特征：由于该文摘系统是面向金融领域的，因此含有金融词汇的句子当然应该设置为较高的权重。在该系统实现时，总结包含一万多个词汇的金融词汇表。(6)中心词特征：文本中权重较高的前 10 个词语，按照标题中的金融词语和非金融词语，确定中心词集合，含有中心词的句子权重大大提高。

本文在结合传统自动摘录的特征组合的基础之上，根据金融领域词汇，对文章中的句子进行权重调整。同时结合标题中的词语和文章词语出现的频率，选出文章中心词，再对文章中的句子进行权重调整。具体算法为：

(1)文本经过分词后，抽取出词性为名词和动词这 2 种实词作为候选特征词，再根据停用词表剔除其中的停用词部分。

根据 TF-IDF 公式对选出的词进行处理，计算出某个特征词在所属句子中的权重 $w_{t,d}$：

$$w_{t,d}=\frac{TF_{t,d}\times\ln(N/DF_t)}{\sqrt{\sum_{t=1}^{m}[TF_{t,d}\times\ln(N/DF_t)]^2}}$$

其中，DF_t 代表文中出现特征词 t 的句子数；N 是文中总句子数；$TF_{t,d}$ 表示 t 在文档 d 中出现的次数。

(2)根据步骤(1)计算每个词在所属句子的权重，扫描金融词汇表，将属于金融词汇的词语增加加权系数 ε_1。

(3)考察文本标题，将标题中候选特征词语集合与金融词汇比较，分别标记为金融词汇标题词 FTW 和非金融词汇标题词 $NFTW$。

(4)根据步骤(2)将全文中权重前 10 位的词语是 FTW 的增加加权系数 ε_2；是 $NFTW$ 的增加加权系数 ε_3，否则不作处理。经过标题词加权的词与标题词的合并集合即为文本中心词集合。

(5)含有指示性词语的句子，设定加权系数 ε_4。

(6)结合句子的位置，以及步骤(3)计算得到的词语权重、步骤(5)计算得到的中心词权重、线索词和指示性短语计算句子的权重，计算公式为：

$$I(s)=\varepsilon_1\times\varepsilon_2\times\varepsilon_3\times\varepsilon_4\times\varepsilon_5\times\varepsilon_6\times\varepsilon_7\times\varepsilon_8\times\frac{\sum_{i=1}^{n}w_{t,d}}{n}$$

其中，n 是句子中的词数；ε_5 是首段加权系数；ε_6 是尾段加权系数；ε_7 是段落首句加权系数；ε_8 是段落尾句加权系数。

4 段落主题划分

在表述某个主题的时候，重点词汇通常限定在围绕该主题涉及到的较小话题范围内，因此，词语具有一定的重复性。可以采用段落相似度方法对文章中的相邻段落按顺序计算其相似度，将文章划分成若干个语义段。在划分主题段落之前，首先使用向量空间模型[6]根据相同词语的有无属性计算相似度划分大类，然后再对每个大类利用向量空间模型根据词语的频度计算段落相似度划分主题。大类的划分可以在很大程度上提高程序的效率。具体的段落相似度算法如下：

(1)设待处理文本共有 n 个段落，为 $P_1,P_2,\cdots,P_n$；文本共有 m 种不同的词语，建立空间向量模型，维度为 m。

(2)对文中所有的段落，计算相邻段落之间词语的共有程度，即相似性。平均相似性=各相邻段落相似性之和/(n−1)。

(3)对于 n−1 个相似性，当发现某个相似性小于平均相似性时，该相似性代表的相邻段落则作为大类的划分界限，后一段落成为新类的开始段落。

(4)经过步骤(3)的大类划分，对于每一个大类，根据词语的频度计算类内每个相邻段落之间的相似性，同样计算每个大类的平均相似性，将低于平均相似性的相邻段落划分开，后一段落成为新主题的开始段落。

5 篇章结构、篇幅分类及文摘句的提取

按照一定的比例提取文本的摘要句，文摘句数=$Nsum\times p$，其中，$Nsum$ 为文本中总句数；p 为限定提取百分比。根据金融领域新闻类和博客类文章的特点，其篇幅结构可能存在这样的情况：在正文中有若干个段落标题，并且每个标题后有相应的几个段落围绕着段落标题进行阐述。第 1 个标题前的段落称之为引语段落，最后一个标题后的段落称之为结束语段落。可以采取这样的规则进行摘要的提取：

(1)抽取每个段落标题作为文摘句，记录下段落标题数 $Ntopic$。

(2)对于引语段落，按照第 2 节中介绍的方法计算句子的权重，选出权重较高的前 n_1 个句子。

(3)对于结束语段落，利用向量空间模型(Vector Space Model, VSM)对其进行段落相似性计算，将紧跟在最后一个段落标题后面的段落相似度较低的段落作为划分界限，从界限开始段落至全文最后一段，从中选取权重较高的前 n_2 个句子。

(4)计算 $m=Ntopic-Nsum$，如果 $m\leqslant 0$，则不做处理；否则计算相应的 n_1 和 n_2。$n_1=m\times$引语段落的句子数/(引语段落的句子数+结束语段落的句子数)；$n_2=m\times$结束语段落的句子数/((引语段落的句子数+结束语段落的句子数)。

对于不满足上述篇章结构特点的文章，则需要根据第 4 节介绍的段落相似度算法，并且按照以下算法进行文摘句的选取：(1)在划分好主题之后，首先计算各个段落主题的权重，公式为：段落主题权重=段落主题内所有句子权重之和/段落主题内的句子数。(2)计算段落主题的总权重，即各主题权重之和。(3)按照比例提取各种主题的相应句数，公式为：$Ntopicparagraph$=段落主题的句子数$\times p\times$段落主题权重/段落主题的总权重。其中，p 为文摘提取的百分数。

6 文摘的平滑处理

文摘的平滑处理具体如下：

(1)冗余主语的处理

将主语相同的相邻句子进行回指词替换。

(2)冗余谓语的处理

将主谓语都相同的相邻句子进行主谓语删除。

(3)冗余连词的处理

例句：但冯某认为，目前恢复审批暂时没有明晰的时间表。这句话前面并没有“虽然”引头的分句，因此，可将“但”删除。同样的类似情况下删除转折性或递进性或表示因果性的连词。

(4)其他情况

例句：目前恢复审批暂时没有明晰的时间表。冯某强调。在文摘生成时，可能会将“冯某强调。”作为文摘候选句，这时需要将其前一句提取出来作为文摘候选句。此情况判断要根据词性标注的结果观察句子是否只有“NR”和“VV”。

7 实验结果与分析

7.1 实验语料及评测方法

本文采用新浪网财经版块的 700 篇文章作为训练语料，300 篇文章作为测试语料。在训练语料中，股票类、理财类、基金类、外汇类、银行类、保险类及黄金贵金属类文章分别 100 篇；文章类型主要为新闻报道、专题评论、专家博客。目前对于自动摘录的评测方法主要分为 2 类：内部评测和外部评测[7]。内部评测测试文摘本身是否与文章要点一致，以及是否包含文章的基本要点；外部评测通过文摘与文章的相似度计算或者文摘在信息检索中所起作用的大小评估文摘好坏的方法。

实验采用现今应用的非常广泛的 ROUGE-2[8]评测方法，通过计算系统产生的文摘和人工文摘 2 元词的共现统计方法计算准确率和召回率并以此评价系统效果。

令 a 为同时出现在自动文摘与人工文摘中的二元词数；b 为自动文摘的二元词数；c 为人工文摘的二元词数。评价指标的计算方式为：准确率为 $P=a/b$；召回率为 $R=a/c$；F_1 值为 $F_1=2\times P\times R/(P+R)$。其中，准确率侧重于文摘精度；召回率侧重于文摘的信息覆盖率；F_1 值表示系统总体性能。

7.2 训练集参数的选取

第 3 节中提到的各个指标的值是通过实验训练得到较好效果(出现次数最多)的一组值。每个参数取值范围为[1,2]，步长为 0.05。

由于语料规模有限，因此可能存在训练语料不充分的问题，本文采取交叉验证策略，将语料随机分为 7 等份，每次选取 100 篇用于训练，最终选取 7 次结果的平均值作为消解结果。表 1 给出实验中表现最好的前 10 个参数集合。

表 1 实验中表现最好的前 10 个参数集合

参数排名	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ε_7	ε_8
1	1.40	1.65	1.50	1.45	1.40	1.35	1.30	1.30
2	1.40	1.65	1.50	1.45	1.45	1.40	1.30	1.40
3	1.30	1.55	1.50	1.55	1.40	1.35	1.45	1.40
4	1.40	1.40	1.55	1.60	1.55	1.40	1.30	1.45
5	1.55	1.30	1.35	1.55	1.40	1.50	1.60	1.55
6	1.35	1.45	1.45	1.55	1.35	1.55	1.40	1.40
7	1.25	1.35	1.40	1.45	1.50	1.35	1.35	1.50
8	1.45	1.35	1.40	1.35	1.50	1.30	1.55	1.35
9	1.35	1.55	1.35	1.30	1.45	1.45	1.40	1.45
10	1.45	1.40	1.60	1.35	1.55	1.30	1.55	1.30

7.3 实验结果

随机选取 300 篇金融类文章，其中，新闻报道、专题评论及专家博客各 100 篇，按照文摘提取比例 20%抽取文摘之后，所得的准确率、召回率和 F_1 值如表 2 所示。

表 2 本文算法的实验结果

文章类型	准确率	召回率	F_1
新闻报道	0.088 474	0.090 736	0.089 590
专题评论	0.081 585	0.084 943	0.083 230
专家博客	0.071 398	0.074 359	0.072 848

由于新闻报道对于文章格式及内容要求标准较严格，因此在篇章结构上的特征表现的比较明显，测试结果的准确率、召回率和 F_1 值较高。由于博客类文章篇章特征决定于作者的个人风格，因此使得其测试结果不是很理想。

表 3 是 300 篇金融类文章的平均评测结果，可以发现，本文算法的评价结果具有明显优势。

表 3 300 篇金融类文章的平均评测结果

算法类型	准确率	召回率	F_1
本文算法	0.080 486	0.083 346	0.081 889
传统 TFIDF 算法	0.068 683	0.070 739	0.069 695

8 结束语

本文提出一种基于指代消解和篇章结构分析的自动摘录算法。在传统基于统计与规则的自动文摘基础上，运用指代消解的思想解决文摘语义不连贯问题，并利用篇章结构的特点选择性地识别出段落标题从而划分段落主题。实验结果表明，该算法提高了自动文摘的准确性。今后将继续针对指代消解进行研究，包括回指词和先行词的识别、概念粒度在指代消解中的应用等，以提高文本语义理解的准确度。

参考文献

[1] 刘 挺, 王开铸. 自动文摘的四种主要方法[J]. 情报学报, 1999, 18(1): 10-19.

[2] 吴 岩, 刘 挺, 王开铸. 中文自动文摘原理与方法探索[J]. 中文信息学报, 1998, 12(2): 8-16.

[3] 傅间莲, 陈群秀. 基于规则和统计的中文自动文摘系统[J]. 中文信息学报, 2006, 20(5): 10-16.

[4] Mitkov R. Robust Pronoun Resolution with Limited Knowledge[C]//Proc. of the 17th International Conference on Computational Linguistics. [S. l.]: ACM Press, 1998.

[5] Lappin S, Herbert J L. An Algorithm for Pronominal Anaphora Resolution[J]. Computational Linguistics, 1994, 20(4): 535-561.

[6] 王 萌, 李春贵, 唐 培, 等. 一种主题句发现的中文自动文摘研究[J]. 计算机工程, 2007, 33(8): 180-181, 189.

[7] 江开忠, 李子成, 顾君忠. 自动文本摘要方法[J]. 计算机工程, 2008, 34(1): 221-223.

[8] Lin Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries[C]//Proc. of ACL Workshop on Text Summarization Branches Out. [S. l.]: BibSonomy Publication, 2004.

编辑 刘 冰

(上接第 169 页)

5 结束语

本文提出一种保持种群多样性的自适应动态粒子群优化算法，通过对典型的抛物线函数——Sphere 函数进行全局优化测试，验证了提高群体多样性有利于提高动态优化算法跟踪全局最优值的能力。将算法应用于群体动画的路径规划中，实现了跟随运动目标的效果，表明算法的有效性。下一步将继续研究提高算法快速适应动态环境的机制，并在更为复杂的环境中进行测试。

参考文献

[1] Shi Y, Eberhart R C. A Modified Particle Swarm Optimizer[C]// Proc. of the IEEE International Conference on Evolutionary Computation. Piscataway, USA: IEEE Press, 1998: 69-73.

[2] Eberhart R C, Shi Y. Tracking and Optimizing Dynamic Systems with Particle Swarms[C]//Proc. of Congress on Evolutionary Computation. Piscataway, USA: IEEE Press, 2001: 94-97.

[3] Carlisle A, Dozier G. Adapting Particle Swarm Optimization to Dynamic Environments[C]//Proc. of International Conference on Artificial Intelligence. Las Vegas, USA: [s. n.], 2000: 429-434.

[4] 高平安, 蔡自兴, 于伶俐. 一种基于多子群的动态优化算法[J]. 中南大学学报: 自然科学版, 2009, 40(3): 731-736.

[5] 焦 巍, 刘光斌. 动态环境下的双子群 PSO 算法[J]. 控制与决策, 2009, 24(7): 1083-1091.

[6] He S, Wu Q H, Saunders J R. Group Search Optimizer: An Optimization Algorithm Inspired by Animal Searching Behavior[J]. IEEE Trans. on Evolutionary Computation, 2009, 13(5): 973-990.

[7] 聂 晶, 刘 弘, 王 琪. 基于粒子群算法的群体动画研究与实现[J]. 计算机工程, 2009, 35(4): 210-211, 214.

编辑 顾逸斐