基于语义统计的中文自动文摘研究

吕静　昝红英

（郑州大学新校区信息工程学院，郑州，450001）

摘　要：自动文摘一直是自然语言处理领域研究的重点和难点。本文在目前的研究状况下，进行了基于语义统计的中文自动文摘研究。主要工作包括：提出一种对 HTML 网页语料进行预处理的方法；利用《同义词词林》，构建概念层次树，在文摘抽取过程中引入了语义信息；通过计算句子重要度,实现了对中文文本的自动摘要和自动索引。试验结果表明，本文提出的方法对多数测试文本都取得了良好效果，优于机械式自动文摘方法。

关键词：同义词词林；自动文摘；概念层次树；主题概念

Chinese automatic abstract based on semantic statistic

LvJing　　ZanHongYing

(The Information Engineering Institute of ZhengZhou University,:ZhengZhou, 450001)

Abstract: The research of automatic abstract is an important and difficult point in the field of natural language processing.After the present research,this paper has presented a Chinese automatic abstract method .The main parts contains several aspects:A pretreating method for web page was developed;Using <<The dictionary of the synonymy words>>,the tree of concept levels was constructed,and the semantic information was incorporated;After computing the sentence importance,the automatic abstract and automatic index for the Chinese texts were achieved. The experimental result showed that the presented method performed better than the mechanical method.

key words: The dictionary of the synonymy words; automatic abstract; the tree of concept levels; topic concept.

1 引言

文摘一般用于检索文献，节省阅读时间，所以文摘要简洁、准确、清晰。本文提出了基于语义统计的自动文摘方法，首先介绍了国内外关于自动文摘的研究状况，然后详细介绍构建基于概念的向量空间模型的几个关键的技术和环节，接下来是测试结果和试验分析，最后将对全文进行总结。

目前，国内外的自动文摘系统主要采用三种方法[1]，基于统计的机械式方法，基于自然语言理解的方法和基于结构的方法。我国对中文自动文摘研究起步较晚，在 20 世纪 90 年代才发展起来。哈尔滨工业大学、复旦大学、北京邮电大学、上海交通大学都分别研制了各自的自动文摘系统。

2 基于统计的自动文摘模型关键技术研究

2.1 对语料（HTML 网页）进行预处理

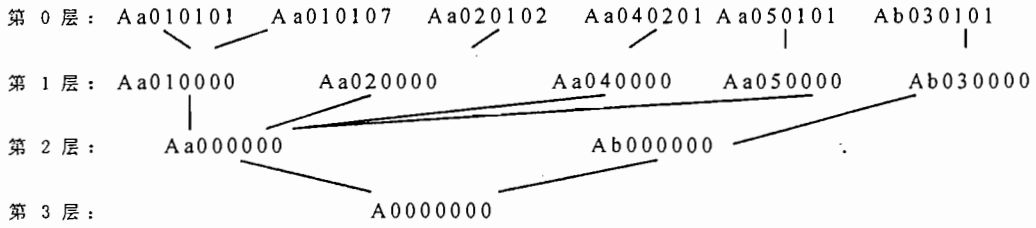
本文的测试需要提前对测试输入的中文 HTML 网页进行预处理。即将文章的标题（Title）提出,将链接、脚本(script)等非文本元素去掉，将 HTML 网页整理成纯文本文件。

作者简介：吕静（1985-），女，河南洛阳人，本科生。E-mail: lvjing080@126.com.

3.2 构造概念层次树

3.2.8 3.2.1 概念层次树的构造规则

此概念层次树按照《同义词词林》[2]构造，分为四层。第 0 层的词语是出现在原文中的词语，第 1 层的词语按照《同义词词林》,从第 0 层抽象而来，依此类推。某段文本按照《词林》标记后，形成以下概念层次树：



图一：概念层次树

Fig.1: the tree of concept levels

3.2.2 概念层次树中词语的参数计算

概念层次树构造好之后，就开始提取主题概念。本算法提出了三个参数[3]：S-频度、T-频度、归纳度 R(C)来衡量一个概念成为主题概念的可能性，下面分别予以介绍。

(1)S-频度：词语 C 的 S-频度为文本中直接表达概念 C 的词语出现的次数。设原文中语义概念为 C 的词的集合为 $\{W_1, W_2 \cdots\cdots W_n\}$,则概念 C 的 S-频度为（其中 $F(W_i)$ 是词语 W_i 在文中出现的频度）：

$$F_S(C) = \sum_{i=1}^{n} F(W_i) \qquad (1)$$

(2)T-频度：设 C 的后代集合为$\{A_1, A_2 \cdots\cdots, A_n\}$,则 C 的 T-频度为(其中 W_T 为折算系数，在此采用 0.9，D(C)是 C 所在的层次，离根节点越近，词语的涵盖能力越强，T-频度就越大。)：

$$F_T(C) = F_S(C) + \sum_{i=1}^{n} (F_S(A_i) * W_T^{D(C)-D(A_i)}) \qquad (2)$$

(3)概念的归纳度 R(C)：如果一个概念的各个子节点对它的 T-频度的贡献比较均衡，那么这个概念的归纳度越强。归纳度的计算公式为：R(C)=1-MAX(所有子节点的 T-频度)/SUM(所有子节点的 T-频度)

(4)概念的选取度：选取度的大小用来衡量概念被选取为主题概念的可能性。其计算公式如下：

$$Sel(C) = (\alpha * (F_S(C) + 1) + \beta * (F_T(C) + 1)) * (\gamma * R(C) + \delta) \qquad (3)$$

其中 α， β， γ， δ 为加权系数，用来调整个参数之间的权重。在此取 α =10, β =0. 25, γ =1, δ =0. 5。

3.3 选取主题概念、计算其重要度、词语与主题概念的相似度

对文本中词语的选取度计算完成之后，选择选取度较大的几个作为主题概念。主题概念的重要度的计算公式为（其中 $F_T(T)$为概念的 T-频度， μ 涉及概念的位置信息，要根据测试集不断调整。）

$$I(T) = \mu * F_T(T) \qquad (5)$$

概念层次树中第 1、2、3 层的词语并不在原文中出现，这些词语与主题概念的相似度定义为二者在《同义词词林》中标号相同的位数。下面举例说明。在《同义词词林》中，"人"的标号为“Aa010101”，“人类”的标号为“Aa010102”。“人”和“人类”的标号前 7 位相同，二者相似度为 7。

3.4 自动索引的创建（关键字的选取）和句子重要度的计算

用文中出现的词语逐个与主题概念进行比较，选取出相似度最大的几个即为文章的关键字。设句子向量 $M(T_1,W_1,S_1;T_2,W_2,S_2;\cdots\cdots;T_n,W_n,S_n)$,其中 T_i 是句子含有的与主题概念相关的词语，W_i 是主题概念的权重，即重要度，S_i 是此词语与主题概念的相似度。句子的重要度要用这些词语的加权平均来计算：

$$I(M)=(\sum_{i=1}^{n}W_i*S_i)/n \qquad (6)$$

计算句子的重要度之后，将其按照重要度排序，选取出重要度较大的句子即为文摘初稿。

3.5 文摘的生成算法

基于以上技术,现提出自动文摘的算法如下:

⑴对输入的 HTML 网页进行预处理，输出纯文本文件。然后用分词算法对文本文件分词，标注词性。

⑵利用《同义词词林》，根据词语之间的语义关系，建立概念层次树。计算各词语的 T-频度、S-频度、归纳度、选取度。根据选取度，并用禁用词表进行过滤，选出文章的主题概念。

⑶计算每个词语与主题概念的相似度，选出文章的关键词，即索引。

⑷计算每个句子的句子向量和句子重要度。根据句子重要度形成文摘。

4 试验与结果分析

4.1 所用语料介绍

本文试验所用语料是项目组人员用网上收集的 HTML 网页建立的语料库，分为科普、经济法律、文学历史、体育旅游等多个类。我们对其进行了人工摘要，机械摘要和索引，以备试验之用。

4.2 试验分析

测试采用准确率来评估：准确率=同时被人工和文摘系统抽取的词语数目/文摘系统抽取的词语数目

表一：测试结果

Tab.1: The test result

准确率 \ 长度 / 类别	4%	6%	12%	12.5%	26%
科普类	0.286	0.250	0.215	0.215	0.238
新闻报道	0.538	0.538	0.673	0.673	0.540
文学历史	0.440	0.430	0.330	0.300	0.316
经济法律	0.550	0.525	0.500	0.562	0.394
体育旅游	0.195	0.264	0.227	0.227	0.180

以上是各类文摘的测试结果，按照文摘长度，其测试准确率如表 1 所示。其中对新闻报道类和经济法律类的测试结果比较理想，而科普类、文学历史和体育旅游类的文摘还需要进一步修改。由于新闻报道类和经济法律类的文章主旨比较突出，中心意思比较集中，用基于语义统计的方法处理时，找到的主题概念能较好地反映整篇文章的主题，所以测试准确率较高。而科普类、文学历史和体育旅游类文章主题比较分散，结构复杂，故准确率不高。另外，文摘的准确率还与文摘长度有关。新闻报道类和经济法律类的文摘长度为 12%左右时，基于词频的算法准确率较高。而其它三类文章的文摘长度要在 26%左右时，语义统计的优势才能显现出来。这主要与输入文本的长度有关。

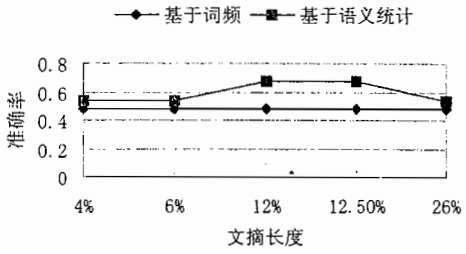


图 2.性能比较图（新闻报道）

Fig2. Comparing Figure(news and report)

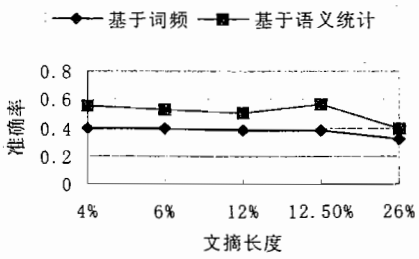


图 3.性能比较图（经济法律）

Fig3. Comparing Figure(Economy and law)

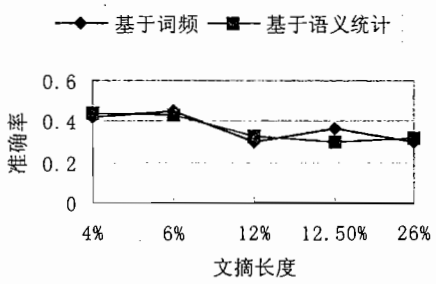


图 4.性能比较图（文学历史）

Fig4. Comparing Figure(literature and history)

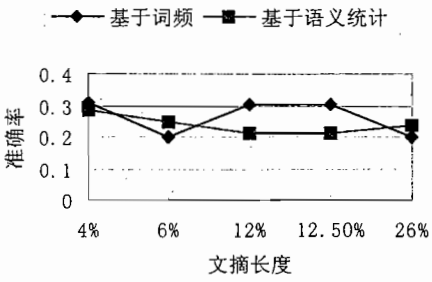


图 5.性能比较图（科普）

Fig5. Comparing Figure(science)

5 结语及进一步工作

本文主要介绍了基于语义统计的自动文摘方法，在自动文摘中结合《同义词词林》引入语义信息，对单纯采用词频统计的方法进行了丰富和发展。本文提出的方法对某些文本的测试取得了良好的效果，但还应适当引入一些自然语言理解、文本分类、信息检索方面的研究成果。

参考文献：

[1]王萌，何婷婷，张伟。基于概念向量空间模型的中文自动文摘系统[J]。《计算机工程与应用》。2005 年 41 卷 1 期。p107-110。

[2]梅家驹，竺一鸣，高蕴琦，等。同义词词林[M]。上海：上海辞书出版社，1983。

[3]季姮，罗振声，万敏，等。基于概念统计和语义层次分析的英文自动文摘研究[J]。《中文信息学报》。2003 年 17 卷 2 期。p14-20。