绿网摘要提取系统算法研究*

龙 珑[1]，邓 伟[2]
(1. 广西师范学院 计算机与信息学院，广西 南宁 530023；
2. 广西肿瘤防治研究所，广西 南宁 530021)

摘 要：随着互联网的普及和发展，传统的文本摘要的提取方法已无法适应绿色网络提供优质内容并过滤不良文本的社会需求。提出通过条件随机场模拟对句子进行注解的方法提取文本摘要。实验证明新方法提取文本的效果有效并可提供更好的过滤不良文本的服务。

关键词：绿色网络；提取信息；不良文本；过滤；条件随机场
中图分类号：TP311 **文献标识码**：A **文章编号**：1674-7720(2013)12-0014-03

Green network the summarization extraction system algorithm

Long Long[1], Deng Wei[2]
(1. Department of Computer Science and Information Technology, Guangxi Teachers Education University, Nanning 530023, China;
2. Guangxi Cancer Institute, Nanning 530021, China)

Abstract: With the rapid growth of internet, the traditional text extraction algorithm can not meet the needs of green network to provide quality content and filter undesirable text . This paper proposed the method of conditional random fields to extract text summarization. At last, the new algorithm was proved speedier and more effective than traditional methods.

Key words: green network; extract information; undesirable text; filter; conditional random fields

目前尚未发现“绿色网络”权威定义，现阶段只能将其理解为可预防网民患上网瘾流行病的计算机网络[1-2]。基于行为分析的绿色网络系统软件的设计目的是为了预防青少年感染不良的网瘾行为。基于行为分析的绿色网络系统中提取文本摘要子系统(下文称绿网摘要提取系统)采用智能的方法浓缩文本信息，从而使基于行为分析的绿色网络系统能快速有效地识别过滤对青少年有不良影响的文本，保留青少年获得感兴趣并对他们身心有益的文本。

在如何快速准确提取文本摘要这个问题上，不少学者进行了大量有价值的研究。Baxendale 提出引入句子位置判断句子重要性的方法提取文本摘要[3]。Luhn 列出高频词并打分，分数高的句子被认定为文本摘要句[4]。AONE C 等提出基于 TF-IDF 朴素贝叶斯模型的算法提取文本摘要[5-7]。KUPIE C J 等通过增加句长方法改进朴素贝叶斯模型的算法提取文本摘要[8]。金立左等提取文本摘要使用基于最大熵模型，增加了先验概率，从而优于所有基于朴素贝叶斯模型的方法[9-11]。

针对文本摘要更新较快和多样性的特点，本文提出通过条件随机场模拟对句子进行注解来达到提取文本摘要的方法。实验表明该方法可有效地提取文本摘要，为绿色网络系统是否过滤该文本提供依据。

1 绿网摘要提取系统摘要主要特征

文本摘要具有三个基本的特征：源自文件、保留文本重要信息、长度短。因此要满足以上特点，很多因素会影响一个句子是否被认为是文本摘要句。影响分为两大类，第一类是句子自身因素，也称单句特征；第二类是文本上下文信息因素，称为关联特征。

1.1 句子自身特征

句子自身特征是指不涉及文本上下文信息也能体现出句子本身的特征。下文列举出几种句子自身因素。

(1)长度特征，由于文本摘要基本不会出现过短或过长的句子，先过滤掉句子中的停用词，然后以词为单位计算目标句子的长度，最终本文算法选取最短和最长的阀值分别为 38 和 6。

* 基金项目：国家创新基金项目(10C26224504901)；国家自然科学基金项目(61163012)；广西自然科学基金项目(2011GXNSFB0180825)

(2)位置特征,位置因素是文本预料的重要特征因素,首句、尾句、段首和段尾是最重要的影响提取文本摘要的几个特征,本文采用了首段、尾句、段首和段尾的位置特征因素。标记词语特征,判断摘要句的标记词语,例如“表示”等,统计表明,约有30%句子含有标记词语,本文算法利用这些词语判断摘要句。

(3)高频词特征,高频词是指出现在目标文本频率较高的句子,词频越高,证明该词语的重要程度越大,所在的句子代表性越强,本文算法在停用词被过滤后,再度量使用高频词。

(4)时间、数字及专有名词特征,文章的焦点基本是命名实体,算法选择句子时决定使用时间、数字以及专有名词。

1.2 上下文关联特征

一个句子是否被绿网摘要提取系统选择为摘要句,除了句子自身的特征影响外,受到上下文关联特征的影响也是很大的。绿网摘要提取系统选择两种基本的关联特征。

(1)与文本标题相似度的关联特征。文本信息包含在标题中,研究发现句子与标题相似度越大,则出现在摘要的可能性越大。

(2)与文本其他句子的相似度的关联特征。实际开发中绿网摘要提取系统把使用该特征看作寻找“高频句”的相似过程,原理与高频词原理相似,建模公式为:

$$f_n=\sum_{n\neq k}\mathrm{Sim}(n,k) \tag{1}$$

其中,$\mathrm{Sim}(n,k)$表示句子n与句子k的相似度,计算公式为:

$$\mathrm{Sim}(n,k)=\frac{|S_n\cap S_k|}{|S_n\cup S_k|} \tag{2}$$

其中,S_n表示组成句子n的词集。句子与上下文句子的相似度在一定程度上能够反应出该句在局部的重要性。

2 绿网摘要提取系统算法的实现

2.1 绿网摘要提取系统的条件随机场模型

条件随机场模型CRFS(Condition Random Fields)是最大熵模型思想的延续,被认为是一种判别模型。在绿网摘要提取系统中,条件随机场模型在给定观察序列的前提下,系统先计算标记序列的概率,然后绿网摘要提取系统再解决序列标注以及标注词性、命名实体标识、分析语块等实际问题。

绿网摘要提取系统的条件随机场模型中,设$\vec{O}$为待标识的观察序列,$\vec{M}$为对应的标识序列,C为观察序列节点集合,则条件概率$p(\vec{M}|\vec{O})$的计算公式为:

$$p(\vec{M}|\vec{O})=\frac{1}{Z(\vec{O})}\prod_{c\in C}\Psi_c(\vec{O}_c,\vec{M}_c) \tag{3}$$

其中,Ψ_c为线性链条件随机场模型中对应最大团的势函数,该函数计算公式为:

$$\Psi_k(\vec{O}_c,\vec{M}_c)=\exp[\sum_{1}^{j}\lambda_n f_n(M_{k-1}\cdot M_k\cdot\vec{O}\cdot k)] \tag{4}$$

$Z(\vec{O})$是以观察序列$\vec{O}$为条件的概率归一化因子,函数计算公式为:

$$Z(\vec{O})=\sum_{\vec{M}\in M}\exp[\sum_{k=1}^{i}\sum_{n=1}^{j}\lambda_n f_n(M_{k-1}\cdot M_k\cdot\vec{O}\cdot k)] \tag{5}$$

其中,$M_{k-1}\cdot M_k\cdot\vec{O}\cdot k$为特征函数,$\lambda_n$是训练中得到的与每个$f_n$相关的权值参数,它反映了特征函数所代表的事件发生的可能性。

2.2 绿网摘要提取系统特征函数

1.1与1.2节中描述了绿网摘要提取系统使用到的特征模板,定义在系统中输入变量为t及输出变量为u,特征函数$f(t,u)$计算公式为:

$$f(t,u)=\begin{cases}1, & (u\equiv c_w)\text{且}(\phi(t)=1)\\ 0, & \text{其他}\end{cases} \tag{6}$$

其中,$\phi(t)$表示绿网摘要提取系统t的特征函数,t成立时,取值为1,则为真。

2.3 绿网摘要提取系统修正因子

如果绿网摘要提取系统选取的摘要句子远小于研究文本中句子数量,则绿网摘要提取系统被选句子的特征出现频率偏低。当绿色摘要提取系统序列标注时,系统目标文本中句子倾向于不被选中,这样绿网摘要提取系统的准确率较高而召回率比较低。从而本文绿网摘要提取系统在条件随机场模型的基础上引入修正因子解决这个难题。计算公式(6)决定判断系统的目标文本中句子s究竟是“在”还是“不在”。

$$\mathrm{tag}(s)=\mathrm{argmax}_{c\in C}\mathrm{Re\ vise}(c)p(c|s) \tag{7}$$

其中$\mathrm{Re\ vise}(c)$为绿网摘要提取系统中类别c的修正因子,$p(c|s)$表示条件随机场模型计算出的类别c的条件概率。从而绿网摘要提取系统目标文本中句子只有两种提取可能性,要么被选中提取,要么不被选中提取,所以系统目标文本中的句子只存在标记“在”和“不在”两种标记,所以$\mathrm{Re\ vise}(\text{在})=1-\mathrm{Re\ vise}(\text{不在})$。

通过大量统计发现,平时使用的训练预料中摘要句子小于或者等于3句的比例高达98.4%,越长的句子被绿网摘要提取系统标记为“不在”的可能性越高。考虑到这一因素,绿网摘要提取系统修正因子的计算公式为:

$$\mathrm{Re\ vise}(x)=\frac{1}{\sqrt{\mathrm{Length}(x)}} \tag{8}$$

其中,x为绿网摘要提取系统目标文本;$\mathrm{Length}(x)$为绿网摘要提取系统目标文本的长度,即系统目标文本所包含的句子数。

3 实验测试结果及分析

3.1 实验测试预料

本文实验的测试数据来源于广西软件测试中心,从搜狐、新浪、网易以及凤凰网4个网站上采集了35 220

篇文本，其中既有不宜青少年阅读的文本，也有适合青少年阅读的文本，将这些平均分为 5 等份，4 份用于训练，1 份用于测试，测试使用交叉验证方法。

3.2 实验测试评测方法

为了更好地评价绿网摘要提取系统测试实验效果，采用准确率、召回率和 F 值 3 个标准指标来衡量，其中 F 值是本次测试最重要的评价指标。绿网摘要提取系统测试实验结果的计算公式为：

$$P=\frac{m}{m+h},\ R=\frac{m}{m+n},\ F=\frac{R\times 2P}{R+P} \tag{9}$$

其中，P 为绿网摘要提取系统测试实验的准确率；R 为绿网摘要提取系统测试实验的召回率；m 为绿网摘要提取系统目标文本摘要中被标记为摘要句的句子数；n 为绿网摘要提取系统目标文本摘要没有被标记为摘要句的句子数量；h 为不在绿网摘要提取系统目标文本摘要中但应该被标记为摘要句的句子数量。

3.3 实验测试设计

本次测试设计为两组。第一组绿网摘要提取系统使用基本的条件随机场模型、最大熵模型和朴素贝叶斯模型进行测试和效果的对比。第二组绿网摘要提取系统使用基本的条件随机场模型和加修正因子的条件随机场模型的测试进行测试和效果对比。

3.4 实验结果及分析

表 1 列出了第一组测试的实验结果。

表 1 3 种模型的系统提取摘要实验结果

系统使用模型	准确率/%	召回率/%	F 值/%
朴素贝叶斯	58.6	62.3	60.4
最大熵	71.3	58.8	64.4
条件随机场	75.2	58.9	66.1

从表 1 的实验结果可以看出，绿网摘要提取系统使用基本条件随机场模型的综合效果（即 F 值）好于系统使用其他两种模型，召回率不如系统使用朴素贝叶斯模型。算法设计小组观察标注的结果发现，当系统选取目标文本过长时，使用条件随机场提取摘要句子分布会过于分散，位于文本中部的句子其位置特征相对分散，导致误判，从而绿网摘要提取系统使用基本条件随机模型提取目标文本的摘要的召回率低一些。算法设计小组对本次测试统计发现，有 54% 的文本超过 10 句，24% 的文本超过 20 句，文本越长，绿网摘要提取系统使用基本条件随机模型提取目标文本的摘要的效果越差。

表 2 列出了第二组测试的实验结果。

从表 2 实验结果可以看出，绿网摘要提取系统使用合适的修正因子条件随机场模型后，召回率提高了 15.4%，综合效果（F 值）也提高了 1.6%，在一定程度上提高了 F 值，取得更好的效果。

绿网摘要提取系统采用增加修正因子的方法改进条件随机模型可以克服目标文本因文本过长所造成的

表 2 系统实验条件随机模型有无修正因子测试结果对比

系统使用修正因子情况	准确率/%	召回率/%	F 值/%
无使用	75.2	58.9	66.1
有使用	72.2	74.3	67.7

影响。从实验效果来看，使用修正因子可以提高提取摘要的效果，今后可重点考虑在算法模型中增加更多因素的修正因子，以提高模型算法的提取效果。

参考文献

[1] 宁葵，龙珑，覃晓，等.绿色网络不良内容语义分析方法研究[J].计算机应用研究，2010，27(12)：4643-4645.

[2] 龙珑，邓伟.绿色网络智能文摘算法研究[J].计算机应用，2012，32(7)：2030-2032.

[3] BAXENDALE P. Machine-made index for technical literature-an experiment[J]. IBM Journal of Research Development, 1958,2(4):354-361.

[4] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research Development,1958,2(2):159-165.

[5] AONE C, OKUROWSKI M E, GORLINSKY J, et al. A trainable summarize with knowledge acquired from robust NLP techniques[C].In Mani, I.and Maybury,M. T., editors, Advances in Automatic Text Summarization, 71-80. MIT Press. 1999.

[6] PANG B,LEE L,VAITHYANTHAN S. Thumbs up? Sentiment classification using machine learning techniques[C]. Processdings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002:79-86.

[7] 何凤英.基于语义理解的中文博文倾向性分析[J].计算机应用，2011，31(8)：2130-2137.

[8] KUPIEC J, PENDERSEN J, CHEN F. A trainable document summarizer[C]. Proceedings of SIGIR'95, 68-73, New York, NY,USA , 1995.

[9] 金立左，袁晓辉，赵一凡，等.二维模糊划分最大熵图像分割算法[J].电子与信息学院，2002，2(8)：1040-1048.

[10] 张龙凯，王厚峰. 文本摘要问题中的句子抽取方法研究[J].中文信息学报，2012，26(2)：97-101.

[11] 屈志毅，李一伟，张延堂，等.一种基于关键重复语义的最大熵文本分类[J].广西师范大学学报(自然科学版)，2007，25(4)：204-207.

(收稿日期：2013-03-26)

作者简介：

龙珑，男，1980 年生，硕士，高级工程师，主要研究方向：人工智能。

邓伟，女，1980 年生，副主任医师，主要研究方向：网瘾防治。