关键词自动标引的最大熵模型应用研究

李素建[1)] 王厚峰[1)] 俞士汶[1)] 辛乘胜[2)]
[1)]北京大学计算语言学研究所 北京 100871)
[2)]人民日报社新闻信息中心 北京 100733)

摘 要 关键词是文档管理、文本聚类/分类、信息检索等领域可利用的重要资源,因此该文提出了利用最大熵模型进行自动标引的技术. 最大熵模型为一个成熟的数学模型,已经应用到计算语言学的各个领域. 然而它的应用非常灵活,针对标引任务和现有资源的实际情况,作者首先建立了最大熵模型的特征集合,然后提出了三种试验方法,并给出了相应的试验结果,最后针对最大熵模型在关键词自动标引任务中的应用做了有益的分析和探讨. 该研究对于关键词标引研究以及最大熵在其他领域中的应用将有所启示.

关键词 关键词标引;关键词抽取;最大熵模型;信息抽取
中图法分类号 TP391

Research on Maximum Entropy Model for Keyword Indexing

LI Su-Jian[1)] WANG Hou-Feng[1)] YU Shi-Wen[1)] XIN Cheng-Sheng[2)]
[1)]*Institute of Computational Linguistics, Peking University, Beijing* 100871)
[2)]*The Information Center of PEOPLE' S DAILY, Beijing* 100733)

Abstract Keywords are very useful to document management, text clustering/classification, information retrieval and so on. Thus, authors propose to use Maximum Entropy (ME) model to conduct automatic keyword indexing. ME model is a mature mathematical model and has been applied to solve many problems. However, it' s flexible how to use the model. Aiming at the indexing task and the resources available, firstly, the useful features for ME model is introduced. Secondly, three tests are given with their experimental results. Lastly, the application of ME model in automatic keyword indexing are analyzed and discussed.

Keywords keyword indexing; keyword extraction; maximum entropy model; information extraction

1 引 言

关键词是文档管理、文本聚类/分类、信息检索等技术可利用的重要资源. 而目前大多文档都不具有关键词,同时手工标引费力费时且主观性较强,不利于下一步的检索工作. 因此关键词自动标引是一项值得研究的技术.

国外对于关键词自动标引的研究起步较早,已经建立了一些实用或试验系统. Turney[1] 设计了系统 GenEx,它将遗传算法和 C4.5 决策树机器学习方法用于关键短语的抽取; Witten[2] 采用朴素贝叶

收稿日期: 2003-07-10; 修改稿收到日期: 2004-06-22. 本课题得到国家"八六三"高技术研究发展计划项目基金(2001AA114210-05)和国家"九七三"重点基础研究发展规划项目基金(G1998030504-01, G1998030507-4)资助.李素建, 女, 1975 年生, 博士, 主要研究领域为计算语言学、信息抽取. E-mail: lisujian@pku.edu.cn. 王厚峰, 男, 1965 年生, 副教授, 主要研究领域为计算语言学、机器翻译. 俞士汶, 男, 1938 年生, 教授, 主要研究领域为自然语言处理、计算语言学. 辛乘胜, 男, 1955 年生, 高级工程师, 主要研究领域为中文信息处理.

斯技术对短语离散的特征值进行训练, 获取模型的权值, 以完成下一步从文档中抽取关键短语的任务. 从国内看, 由于汉语语言本身的特点, 没有显式的词边界, 为主题自动标引任务又增加了一定的难度, 使用最多的一种方法是基于 PAT Tree 结构获取新词, 并采用互信息等一些统计方法对文档的关键词进行标引[3], 但获取候选词选用的 PAT Tree, 它的建立用计算机实现需要大量的空间消耗, 因此还需要进一步深入研究.

最大熵方法是当前自然语言处理领域最为盛行的一种方法, 在 Conll-2003[4] 的 NER(命名实体识别)比赛中, 16 个参赛小组中的前三名都提到了该方法. 此外, 最大熵方法还有效地应用到词性标注、歧义消解、边界识别、浅层分析等领域[5~8]. 这也说明该方法易行且有效, 但目前还未见报道把最大熵方法用于关键词自动标引的工作中. 因此本文针对关键词标引任务, 在最大熵模型概率计算的基础上, 探索了多种运用最大熵模型的试验方法.

本文第 2 节介绍了关键词自动标引的任务以及需要解决的问题; 第 3 节结合关键词标引任务回顾了最大熵模型, 并描述了模型中使用的特征集合; 第 4 节详细介绍了基于最大熵模型的三种试验方法; 第 5 节给出了试验结果, 并进行比较和分析; 最后对全文进行了总结.

2　关键词自动标引任务简介

关键词自动标引是根据文档的主题内容, 借助计算机处理技术, 自动从文档中直接抽取关键词作为标引词. 因此有人也把关键词自动标引称作关键词抽取技术. 这里的关键词不局限于一般的词的概念, 也可能为一个短语, 由多个词语构成.

实际上, 一个文档可以表示成一个广义集合①, 集合中的每个元素为具有出现频数、出现位置等属性的短语. 因此, 关键词标引的任务就是判断这个集合中哪些短语是关键词和哪些是非关键词. 这样就需要解决两个问题, 第一个是如何将文档表示成广义集合, 即从中提取出哪些短语作为关键词候选项, 如何提取; 第二个问题是怎样判断候选项是否是关键词, 其依据是什么.

一篇文档就是一个字符串序列, 如果把文档中所有可能的短语抽取出来, 这个数目是非常庞大的, 而且很多是不必要的, 例如一些虚词和由虚词组成的短语不能作为关键词. 这里, 我们利用一些语言学工具从文中选出在一定程度上反映了文章主题内容的有意义的短语. 由于文档中重要的内容经常会重复出现, 因此首先由串频统计工具[9] 从文中获得出现一定频数的字串, 再根据词性切分标注器、浅层分析器等工具以及语言学知识把没有意义的和不可能作为关键词的字串过滤掉, 得到一个关键词候选项的集合, 这里不再详细介绍候选项获取的过程, 可参考 Li 的文章[10].

当前关键词候选集合中的每一项都在一定程度上反映了文章的内容, 但其反应主题内容还存在着程度大小的问题. 而获取文档的关键词就是选出最能反映文章主题内容的那些候选项. 因此要计算每一候选项反应主题内容的程度大小, 最大熵模型就是我们计算这个程度大小以获取关键词的基础.

3　最大熵模型

最大熵模型是一个比较成熟的数学模型, 适合于估计事件的概率分布. 最大熵框架的计算模型不依赖语言模型, 独立于特定的任务. 这里我们再简单回顾一下最大熵框架的原理. 进行关键词标引, 我们选取训练数据时, 以每一个字串作为一个事件. 假设有一个样本集合为 $\{(ck_1, y_1), (ck_2, y_2), \cdots, (ck_N, y_N)\}$, 每一个 $ck_i (1 \leq i \leq N)$ 表示一个进入最大熵模型进行概率估计的候选关键词(candidate keyword), $y_i (1 \leq i \leq N)$ 表示该候选项被标引的结果, 该结果属于集合 {YES, NO}, YES 表示是关键词, NO 表示不是关键词. 利用最大熵框架模型得出在特征限制下最优的概率分布, 即概率值 $p(y \mid ck)$. 根据最大熵原理, 概率值 $p(y \mid ck)$ 的取值符合下面的指数模型:

$$\begin{cases} p(y \mid ck) = Z_\lambda(ck) \exp\left(\sum_i \lambda_i f_i(ck, y) \right) \\ Z_\lambda(ck) = 1 \Big/ \sum_y \exp\left(\sum_i \lambda_i f_i(ck, y) \right) \end{cases} \tag{1}$$

这里 f_i 表示候选项所具有的可能特征, 它是一个二值函数, 描述某一个特定的事实. λ_i 指示了特征 f_i 对于模型的重要程度. $Z(ck)$ 是一个范化常数. 公式(1)使模型由求概率值转化为求参数值 λ_i, 我们采用的方法是 Darroch 和 Ratcliff[11] 的通用迭代缩放算法(Generalized Iterative Scaling, GIS), 用来得到

① 张学文. 组成论: 广义集合和复杂度定律. http://entropy.com.cn, 2001

具有最大熵分布的所有参数值 λ_i. 获得参数值后, 就可以得到事件的概率分布, 这里表示获得候选项是关键词或者不是关键词的概率分布.

最大熵模型中, 特征集合的选取是一个非常重要的问题. 这里我们介绍一下关键词自动标引任务中最大熵模型用到的可能特征. 这时, 我们把关键词标引看作从候选项集合中挑选关键词, 一般考虑以下的一些可能因素, 如候选项的长度、其在文档中的出现次数、是否包括在特殊符号(如引号或书名号内)、在文中首次出现的位置、是否为命名实体等. 根据这些因素, 我们可以建立一个特征模板, 并根据训练数据定义每个模板的取值范围. 如表 1, 模板 1 ~ 7 可以看作是影响关键词标引的特征模板, 模板 8 是一个特殊的模板, 表示标引结果.

表 1 特征模板及取值范围

模板号	模板意义	模板	取值范围
1	长度	LEN	Int{2 ~ 10}, MORE{≥10}
2	出现频数	FREQ	Int{1 ~ 19}, MORE{≥20}
3	短语类型	SYNTAG	{NP, VP, OTHER}
4	是否在特殊符号(引号或书名号)中	IN-PUNC	{Quotes(“ ”), Brackets(《》), NONE}
5	首次出现位置	POSITION	{Title, First Paragraph, Last Paragraph, OTHER}
6	文档类型	STYLE	{National News, Sport International News, OTHER}
7	命名实体类型	NE-TYPE	{PERSON, PLACE, COUNTRY, ORGANIZATION, OTHER}
8	标引结果	DEFAULT	{YES, NO}

当模板函数取特定值时, 则该模板被实例化, 得到具体的特征. 取模板 1 ~ 7 中的任一个模板, 确定该模板的取值, 并结合当前标引结果的值(即模板 DEFAULT 的取值). 当两个模板的取值确定后就可以产生一个特征. 下面我们对特征的格式进行一下定义, 每一个特征由三部分构成:

(1)第一部分是下划线“-”前的部分, 为特征模板, 如 LEN、FREQ 等, 表示标引时要考虑的影响因素;

(2)第二部分是下划线和等号中间的部分, 例如 2、MORE 等, 表示该特征模板的取值;

(3)第三部分为模板 DEFAULT 的取值, 即表示标引结果, 是否为关键词.

在得到一个特征后, 该特征可以表示为二值特征函数的形式.

例如, 由特征模板 1 可以得到一个特征 $LEN\text{-}MORE = NO$, 表示为二值特征函数为

$$f_j(ck, y) = \begin{cases} 1, & LEN(ck) = MORE \text{ 且 } y = NO \\ 0, & \text{其它} \end{cases} \quad (2)$$

$LEN(ck)$可以看作模板 LEN 的函数表示, y 是模板 DEFAULT 的取值, 表示当前输出的结果. 该特征函数表示如果候选项的长度大于或等于 10, 并且其不为关键词, 则函数值为 1, 否则为 0.

确定特征集合后, 对训练数据进行参数估计. 测试阶段中, 首先从新文档中获取候选项集合, 对每一项分别依据其特征值计算成为关键词或非关键词的概率值, 最后由这些概率值来确定文档的关键词集合. 下面将分别采用不同的试验方法获取文档的关键词.

4 试验方法

前面介绍了关键词自动标引的任务以及最大熵模型的本质, 即根据已有的事例获得最优的概率分布. 如何更好地利用最大熵模型完成这个任务, 我们做了一些有益的探索, 提出了三种以最大熵模型为基础的试验方法: 分类试验、正例试验、打分试验. 下面对各种方法的具体实现过程分别进行介绍.

4.1 分类试验

自然语言处理中很多问题如词性标注、组块分析等任务采用最大熵模型的基础是把这些问题看作分类问题, 通过概率值的计算和比较, 选择具有最大可能发生的类别. 而关键词标引的问题, 我们同样也可以看作一个分类问题, 计算一个候选项是或不是关键词的概率值, 并比较两者大小, 概率值大的标引结果则是最终的标引结果. 这是我们进行分类试验的理论基础.

由表 1 中的特征模板进行实例化, 可以得到一个特征集合. 训练语料中的文档已经手工标注关键词, 同时由自动生成程序为每一个文档生成一个关键词候选项集合. 每一个候选项可以看作一个事件, 根据特征函数集合, 建立一个特征向量. 由于每一个特征函数都是二值函数, 因此每一维为 0 或 1. 如果候选项属于手工标引关键词集合, 则为正例, 否则为反例. 对于第三部分构成为“NO”的特征函数, 正例

在该维的取值均为 0. 同样, 对于第三部分构成为“YES”的特征函数, 反例在该维的取值也均为 0. 这时, 输入特征向量集合, 采用 GIS 算法获得最大熵模型的参数.

对于新文档, 首先自动获得一个关键词候选集合, 然后对于每一个候选项分别假设其为关键词(即模板 DEFAULT 取值 YES), 并根据该候选项的特征获得一特征向量. 把训练得到的参数值代入公式(1), 可以估计该候选项为关键词的概率值, 然后假设该候选项不为关键词(即模板 DEFAULT 取值为 NO), 同样可以获得一特征向量, 那么可以得到其为反例的概率值. 候选项为正例和反例的概率值分别设为 *ProbP* 和 *ProbN*, 比较其大小, 若 *ProbP*≫ *ProbN*, 则为关键词, 否则不为关键词.

4.2　正例试验

由于在关键词标引任务中, 对于候选项成为关键词具有贡献的特征只有第三部分构成为 YES 的特征, 也就是特征模板 DEFAULT 取值 YES. 因此, 我们只需考虑得到这些特征函数的参数值, 并估计每个候选项成为关键词的概率值. 值越大, 其成为关键词的可能性也就越大.

首先该试验建立特征集合, 所有特征均由特征模板 DEFAULT 取值 YES 构成. 同时只收集正例, 也就是只对手工标注的关键词根据特征集合生成一个特征向量, 然后由 GIS 算法获得各特征的参数值, 参数值直接反映了相应特征对于成为关键词的贡献.

测试时, 对于一篇文档生成的所有候选项, 利用公式(1)分别得到每一项成为关键词的概率值, 然后根据文档的长短和内容选取概率值最大的几项作为文档的关键词.

4.3　打分试验

该试验结合了分类试验和正例试验的方法, 特征和训练语料的选取如分类试验, 特征模板 DEFAULT 可以取值 YES 和 NO. 训练语料也包括了正例和反例. 对于训练语料中的所有候选项分别建立特征向量, 然后用 GIS 算法得到每一个特征的参数值.

测试时, 首先生成一篇文档的候选项, 对于每个候选项首先假设其成为关键词(模板 DEFAULT 取值 YES), 根据公式(1)计算概率(*ProbP*), 并假设其不为关键词(模板 DEFAULT 取值 NO)并获得概率(*ProbN*). 然后如公式(3), 两概率相除, 为该候选项(*ck*)赋一个分值 *score*(*ck*).

$$score(ck) = \frac{p(y = \mathrm{YES} \mid ck)}{p(y = \mathrm{NO} \mid ck)} \tag{3}$$

在选取关键词时, 借鉴了正例试验的思想, 根据文档的长短和内容选取分值高的几项作为文档的关键词.

5　试验结果及分析

我们从 100 篇新闻文档中抽取了 2260 个词或短语作为关键词候选项, 其中包括了 2015 个反例和 245 个正例, 反例表示候选项不是关键词, 正例表示候选项为关键词(手工标引). 用精确率、召回率、综合指标 $F_{\beta=1}$ 来评测标引的结果, 其定义如下:

$$\begin{aligned} \text{精确率} &= \frac{\text{自动标引正确的关键词数目}}{\text{自动标引的关键词数目}} \\ \text{召回率} &= \frac{\text{自动标引正确的关键词数目}}{\text{手工标引的关键词数目}} \\ F_{\beta=1} &= \frac{2\times \text{精确率}\times \text{召回率}}{\text{精确率}+\text{召回率}} \end{aligned} \tag{4}$$

在试验评测中, 召回率和精确率是一对矛盾的指标, 当自动标引过程中选择关键词的数目增大时, 一般召回率增大, 精确率减小; 反之当选择关键词的数目减少时, 则召回率减小, 精确率增大. 由于候选项都是从原文中抽取得到的, 在一定程度上反映了文章的主题内容, 所以在保持精确率和召回率相对平衡的基础上, 尽量提高关键词标引的召回率. 其中我们以国内新闻为例, 说明指标随着关键词的数目发生变化.

采用打分试验方法, 表 2 列出了关键词自动标引数目对于各指标的影响. 表中的数目比例表示关键词数目的选择占文档中自动生成候选项数目的比例, 因为候选项的多少往往与文档内容和长短紧密相关. 从表中可以看出, 关键词数目占候选项数目的比例约为 1/5 时, 其综合指标明显低于其他比例的综合指标. 当取值为 1/10 和 1/7 时, 综合指标 $F_{\beta=1}$ 值相差不多, 但 1/7 比例的召回率要比 1/10 要高 4~5 个百分点, 因此这种情况下, 我们认为关键词数目取值比例约为 1/7 时, 系统性能最为理想. 对于不同类型的文档, 关键词数目取值比例也要适当调整.

表 **2**　关键词数目与指标的关系

	召回率	精确率	$F_{\beta=1}$
取值比例约为 1/10	0.369	0.423	0.384
取值比例约为 1/7	0.405	0.359	0.381
取值比例约为 1/5	0.423	0.276	0.334

选取 96 篇新闻文档并手工标注关键词,作为测试语料.其中 44 篇国内新闻,36 篇国际新闻和 16 篇体育新闻.表 3 列出了三种试验方法进行自动标引得到的评测结果,同时对照了 Li 的文章[10]中的经验打分的试验结果.

表 3 各种试验方法的评测结果

	召回率				精确率			
	分类试验	正例试验	打分试验	Li 2003	分类试验	正例试验	打分试验	Li 2003
国内新闻(44)	0.151	0.262	0.445	0.452	0.726	0.170	0.397	0.401
国际新闻(36)	0.240	0.323	0.594	0.644	0.761	0.279	0.548	0.594
体育新闻(16)	0.161	0.284	0.566	0.629	0.626	0.239	0.439	0.482
平均值	0.186	0.289	0.521	0.554	0.722	0.222	0.461	0.487

在表 2 中,分类试验的效果最差,其召回率几乎接近系统的底线(baseline).系统的底线指:根据专名词典,直接从文中匹配并抽取出一些专有名词作为关键词.由于分类试验从候选项中获取其他关键词的能力很弱,这里的精确率主要体现了根据专名词典抽取得到关键词的准确度.所以单独一个指标不能说明一个系统的性能好坏.其次是正例试验,其结果中召回率比分类试验平均提高了 10 个百分点.打分试验的结果最为理想,其结果可以达到当前标引任务的水平.

下面我们结合实际的训练语料,对最大熵模型在标引任务中三种应用的试验结果进行分析.分类试验是把关键词标引的问题转化为分类问题来解决,首先观察训练数据的分布,因为两类事例选取的不平衡性,其中反例和正例的个数之比几乎为 9∶1,在这种情况下,对于候选项成为反例的特征通常要比候选项成为正例的特征贡献要大.因此对于新文档的候选项进行概率计算时,反例概率 *ProbN* 一般大于正例概率 *ProbP*,所以几乎没有候选项被分类为正例,即被标为关键词,这也是该方法召回率很低的原因.该方法是根据类别概率的大小确定是否为关键词,因此生成的关键词数目也是唯一的,不能够进行调整.该方法显然不适合用于关键词标引的工作中.通过对该方法的分析,我们总结出,采用最大熵模型要结合当前任务和实际的训练语料,不能都作为简单的分类问题处理,要具体问题具体分析.

对于正例试验,在选择训练语料时我们只考虑了正例,相应也只考虑和正例相关的特征.这样相对于分类试验,我们排除了很多噪音对于候选项成为关键词的反面影响,通过参数估计,特征参数也在一定程度上反映特征对于成为关键词的贡献大小,同时通过概率计算获取概率值较大的几项作为关键词.但由于模型是针对正例进行训练,因此只考虑特征对成为关键词的正面作用,没有考虑其反面作用.例如,假设所有正例中长度为 2 的项很多,那么在该试验中特征函数包含"*LEN*-2=YES"的候选项成为关键词的概率值就会很大,但是如果综合考虑所有正例和反例,就会发现特征函数包含"*LEN*-2=NO"的候选项数目要大于具有上面特征函数的候选项数目,因此特征函数包含"*LEN*-2=YES"不一定有助于成为关键词.忽略了反例事例,就会把一些影响事例成为非关键词的因素忽略掉,这是造成该试验方法标引不准确的一个重要因素.

打分试验方法综合考虑了最大熵模型中影响成为关键词的特征以及影响成为非关键词的特征,为每个候选项进行打分,分值与为正例的概率值成正比,与为反例的概率值成反比.也可以看作,对成为关键词具有贡献的特征增大了分值,对成为非关键词具有贡献的特征减小了分值.因此这种方法应用到关键词标引任务中还是合理的.但是其结果比 Li 的文章[10]中的经验打分结果要差一些,原因在于目前的训练数据少,而且训练数据为人工标引的结果,带有一定的主观性,所以在参数估计时会出现一些不一致的现象.如果克服了主观性的因素,那么最大熵方法还是最有潜力的,它的参数选取都有严格的数学推导,不像经验打分一样还需要反复试验参数值,不具有理论依据.

6 结 论

本文通过各种成熟的语言学工具首先从文档中获取关键词候选项,并建立了一个特征集合,利用最大熵模型根据丰富的语言特征来判断候选项是否可以作为文档的关键词.因此,可以看出最大熵模型的优点在于可以灵活地选择各种特征,结合大量的特征到模型中去.

文中对各种基于最大熵模型的关键词自动标引方法做了试验和探讨其中包括分类试验、正例试验、

打分试验.由于训练数据中正例数目和反例数目的比例相差过大,总结出分类试验不适合用于关键词标引任务中;正例试验则没有充分考虑影响候选项成为非关键词的特征,所以效果也不是很好;而打分试验综合考虑了为正例和为反例做贡献的特征,更适合用于关键词标引任务中.因此,我们知道最大熵模型本身提供了如何计算最优的概率分布,对它的应用则可以很灵活,要根据实际的工作和资源,选用适当的方法来解决问题.

尽管最大熵模型在关键词标引中的应用并不是非常理想,没有体现出比其他方法的优越性.很大原因是因为训练数据数量的限制,从而影响了特征的选择,以及估计特征参数时有欠准确.目前我们只有2000 多个事例,约 100 个特征,就可以达到当前标引的水平,说明最大熵模型本身还有很大潜力.这样我们下一步的工作,要在现有标引模型的基础上,自动生成文档的关键词,并加以人工校对,不断积累训练语料,改善最大熵模型,提高关键词标引的性能.

致　谢　北京大学计算语言学研究所的段慧明老师、人民日报社新闻信息中心的盛成厚老师对本工作给予了大量帮助,叶嘉明等同学在程序开发等方面做了很多工作,北京大学计算语言研究所其他同仁也给予了大力支持,以及评审老师也提出了中肯意见,在此一并表示感谢.

参　考　文　献

1 Turney P. D.. Learning to extract keyphrases from text. National Research Council, Canada, NRC Technical Report ERB-1057, 1999

2 Witten I. H., Paynter G. W., Frank E., Gutwin C., Nevill-Manning C. G.. KEA: Practical automatic keyphrase extraction. In: Proceedings of the 4th ACM conference on Digital libraries, Berkeley, California, US, 1999, 254~256

3 Yang Wen-Feng. Chinese keyword extraction based on max-duplicated strings of the documents. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 2002, 439~440

4 Tjong E. F., Sang K., Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-independent named entity recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada, 2003, 142~147

5 Ratnaparkhi A.. A maximum entropy model for part-of-speech tagging. In: Proceeding of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, USA, 1996, 133~141

6 Ratnaparkhi A.. A simple introduction to maximum entropy models for natural language processing. Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, USA: Technical Report 97-08, 1997

7 Wojciech S., Brants T.. A maximum entropy partial parser for unrestricted text. In: Proceedings of the 6th Workshop on Very Large Corpora, Montreal, Canada, 1998, 143~151

8 Koeling R.. Chunking with maximum entropy models. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000, 139~141

9 Liu Ting, Wu Yan, Wang Kai-Zhu. An Chinese word automatic segmentation system based on string frequency statistics combined with word matching. Journal of Chinese Information Processing, 1998, 12(1): 17~25

10 Li Su-Jian, Wang Hou-Feng, Yu Shi-Wen, Xin Cheng-Sheng. News-oriented automatic Chinese keyword indexing. In: Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, 2003, 92~97

11 Darroch J. N., Ratcliff D.. Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics, 1972, 43(5): 1470~1480

LI Su-Jian, born in 1975, postdoctor. Her research interests include natural language processing, computational linguistics, information extraction.

WANG Hou-Feng, born in 1965, associate professor. His research interests include computational linguistics, machine translation.

YU Shi-Wen, born in 1938, professor. His research interests include natural language processing, computational linguistics.

XIN Cheng-Sheng, born in 1955, senior engineer. His research interest focuses on Chinese information processing.

Background

This project, "Research and system development of content indexing for news document", is supported by the National High Technology Research and Development Program (863 Program) under grant No. 2001AA114210-05 and National Basic Research Program of China (973 Program) under grant No. G1998030504-01 and G1998030507-4. Its goal is to develop a new technique that can manage large collections of news documents effectively. Authors have implemented one prototype system as a testing bed. This paper adopts the technique of maximum entropy for keyword indexing.