工程与应用

基于 N-gram 统计模型的搜索引擎中文纠错

陈智鹏,吕玉琴,刘华生,刘 刚,屠 辉

(北京邮电大学 电子工程学院,北京 100876)

摘 要:搜索引擎中的关键词纠错是提高检索效率的一项重要辅助功能。提出了一种完全通过分析上下文统计信息的方法,根据中文语言的特点,在建立 N-gram 统计模型并分析比较的基础上,再通过计算 TF/IDF 的权重来获得最优的纠错结果,最后通过实验验证了该方法实现了搜索引擎中对输入关键词的自动检查和纠错。

关键词:搜索引擎;输入纠错;N-gram 模型;TF/IDF

中图分类号:TP393 **文献标识码**:A **文章编号**:1673-5692(2009)03-323-04

Chinese Spelling Correction in Search Engines Based on N-gram Model

CHEN Zhi-peng, LV Yu-qin, LIU Hua-sheng, LIU Gang, TU Hui

(Beijing University of Posts and Telecommunications School of Electronic Engineering, Beijing 100876, China)

Abstract: Key words spelling correction plays an important part in the improvement of efficiency in a search engine. In this article, a method that analyzes only the context-sensitive statistics is discussed. According to the characteristics of the Chinese language, this method is based on the establishment of N-grams model and the analysis and comparison of it, and it involves the calculation of the TF/IDF weights to obtain the best error correction. This correction model is tested in actual practice and is proved effective.

Key words: search engine; spelling correction; N-grams model; TF/IDF weight

0 引 言

随着信息社会的发展,对搜索引擎的要求越来越高,而对输入关键词进行自动纠错是搜索引擎所要求的基本应用之一。

搜索引擎的输入自动纠错功能[1]是指,用户在输入关键词进行搜索的时候,如果搜索引擎在返回结果中计算出与此关键词相似的另一形式(如词组中出现同音不同字,或者某一错别字现象)得到大量的搜索结果,用户将会在搜索结果页面看到系统推测提供的关键词项。

搜索引擎中的自动纠错系统结合了计算语言学的信息检索和文本自动勘校两方面的应用,也包括了信息论(information theory)在自然语言处理技术中的应用。

在英文的文本自动勘校[2,3]中,因为英文中每个词之间有空格,不用考虑分词问题,所以只需要对每个单词进行拼写检查(spelling detection),常用的方法是利用编辑距离(levenshtein distance)来确定词与词之间的相似程度,另外考虑每个词在文本中的统计信息来最终判断错误拼写。而对于中文,考虑到汉语语言的特殊性,首先要对文本进行切割分词,然后再进行错误检查和校正,常见的方法包括基于字典和基于文本统计信息两种。基于字典是要求建立一个庞大的字典库,需要一定的维护代价,而随着网络和自然语言的飞速发展,仅仅依靠不断扩大的词典的收录规模来进行纠错越来越难以满足搜索

收稿日期:2008-11-10 修订日期:2009-01-12

引擎所要求的效率[1]。

在此提出一种将完全基于上下文统计信息的文本校对方法应用于中文信息检索领域中的自动输入纠错。

1 建立统计模型及优化筛选

1.1 基于 N-gram 的语言模型

中文信息检索系统中,必不可少的要对进行检索的文本进行分词,这就需要首先建立其统计语言模型(statistical language models)。N-gram 是最为常用的统计语言模型, Mays[4]在英文的自动校对中应用了 Trigram 模型,施得胜[5]和张仰森[6]在中文的文本校对中使用了 Bigram 模型,马金山等[7]则提出了利用 trigram 和上下文依存分析来进行中文自动查错,取得了一定的效果。

从统计角度看,自然语言中的一个句子 s 都是由一连串特定顺序排列的词 $w_1,w_2,w_3,\cdots,w_n$ 构成,s 出现的概率 $P(s)$ 为

$$P(s)=p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\cdots p(w_l|w_1\cdots w_{l-1})=\prod_{i=1}^{l}p(w_i|w_1\cdots w_{i-1}) \quad (1)$$

可以认为对于每个词 w_i,它的出现概率取决于它前面所有词。从计算上来看,各种可能性太多,计算量太大,无法实现。因此我们假定任意词 w_i 的出现概率只同它前面的 $n-1$ 个词有关,即为 N-gram 模型

$$p(w_i|w_{i-n+1}\cdots w_{i-1}) \quad (2)$$

N-gram 模型又被称为一阶马尔科夫链。

对于 N 值的选择,由真实的语言和其模型的交叉熵来决定。语言 $L=(X_i)\sim q(x)$ 与其模型 p 的交叉熵定义为

$$H(L,p)=-\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n}q(x_1^n)\log p(x_1^n) \quad (3)$$

式中,x_1^n 为随机变量$(x_1,x_2,\cdots x_n)$,表示长度为 n 的自然语言序列,其中每个随机变量 $x_i(i=1,2,\cdots,n)$ 代表自然语言序列上的一个汉语语言单位词。x_i 可在其所代表的词集 $\mathbf{X}$ 中取值。自然语言序列可被视为离散的平稳有记忆信源。

假设这种语言 L 是"理想"的,即上文所述中提到的,即 n 趋于无穷大时,其全部"单词"的概率和为 1。那么可以假定语言 L 是稳态(stationary)随机过程,信源是各态遍历的,根据 Shannon-McMillan-Breiman 定理,交叉熵可由

$$H(L,p)=-\lim_{n\to\infty}\frac{1}{n}\log p(x_1^n) \quad (4)$$

求出,$x_i\in\mathbf{W}$,式(4)可由统计语料库近似求出。

对于 N-gram 语言模型中,可以计算出句子的概率为

$$P(s)=\prod_{i=1}^{m+1}P(w_i|w_{i-n+1}^{i-1}) \quad (5)$$

假定测试语料 T 由 l_T 个句子构成$(t_1,\cdots t_{l_T})$,则整个测试集的概率为

$$P(T)=\prod_{i=1}^{lT}P(t_i) \quad (6)$$

模型 $P(w_i|w_{i-n+1}^{i-1})$ 对于测试语料的交叉熵

$$H_p(T)=-\frac{1}{W_T}\log_2 P(T) \quad (7)$$

其中 W_T 是测试文本 T 的词数。模型 P 的困惑度 $PP_P(T)$ 定义为

$$PP_P(T)=2^{H_P(T)} \quad (8)$$

由上述推理得出,对于理想的 N-gram 语言模型而言,其交叉熵越小,困惑度也越小,语言处理的能力就越好,也就是语言模型与真实语言越接近。欲使交叉熵和困惑度变小,则需要选择较大的 N 值,就可以提供更多的语境信息,更具区别性,但是计算代价会过大,训练语料需要多,考虑到现实机器的处理能力,语言自身特点等因素,不宜选择过大的 N 值。

在《现代汉语语法信息词典》录入的词语中,单字词占 7.36%,两字词占 63.3%,三字词占15.33%[8],也就是在汉语中单字词、双字词和三字词总共占了汉语语言样本中 99.2% 的比重,所以在此提出的应用场景中,将为所提供的中文检索文本和输入关键词(句子)建立 n 分别为 1,2,3 的语言模型,即合并了 unigram($n=1$)模型,bigram($n=2$)模型和 trigram($n=3$)模型的统计语言模型。

1.2 输入关键词的分析及统计信息比较

用户向搜索引擎提交长度为 l 的关键词后,如果在索引中查询返回结果不为空,则认为用户输入的关键词用检查和纠错。如果搜索返回结果为空,则认为该关键词有可能是用户输入错误,系统会将该关键词根据 N-gram 模型切分,设查询的关键词长度为 l,则切分之后的词的个数为:$3(l-1)$。

另外,根据 N-gram 语言模型,对于全部语言样

本进行切分,并且建立索引,自动的输入纠错功能将搜索关键词切分后得到的 $3(l-1)$ 个词分别在此索引进行搜索。

在合并了分别用 unigram,bigram 和 trigram 分词的语言样本的索引后,$3(l-1)$ 个词分别返回其在原语言样本中出现的位置,最终比较切分后的子关键词出现位置的次数,返回出现次数最多的位置,设此次数为 $n[0\leqslant n<3(l-1)]$,并根据在此位置出现的子关键词的长度,返回长度为 l 的短语。

例如:输入关键词"北京奥运会"搜索返回为空,长度 $l=5$,则对其按照 unigram,bigram 和 trigram 切分为 12 个子关键词在原文中搜索,返回结果出现的位置次数最多为 9,即 $n=9$,而这些位置出现的词为"北京奥运会",即最终返回"北京奥运会"这个长度同为 $l=5$ 的短语。

2　用 TF/IDF 方法计算搜索关键词的权重

词频率 - 逆向文档频率(TF/IDF)是来自于信息论中相对熵(kullback-leibler divergence)的应用。相对熵通常用来评价两个随机分布 $p(x)$ 和 $q(x)$ 的差距

$$D(p\parallel q)=\sum_{x\in\mathbf{X}}p(x)\log\frac{p(x)}{q(x)} \tag{9}$$

当两个随机分布相同时,其相对熵为 0;当两个随机分布的差别增加时,则相对熵也增加。

词频率(TF,term frequency)TF_{ij} 是指词 T_j 在文本 D_i 中出现的频率,是关于查询词语和相应文本的相关性的度量;文档频率(DF,document frequency)DF_j 是指词 T_j 出现过的文本总数,逆向文档频率指数(IDF,inverse document frequency)的公式为

$$IDF_j=\log\frac{N}{DF_j} \tag{10}$$

其中 N 代表所有文档的总数。

计算 IDF 的值是信息检索中使用最多的衡量关键词权重的方法,因为在一定量的文本中,如果一个词的文档频率(DF)越小,则说明该词预测主题能力越强,权重就越大,反之,权重就越小。

以上求出两值相乘结果,即代表修正过的关键词 T_j 在文档 D 中的权重,如公式描述

$$W_{ij}=TF_{ij}\times IDF_j \tag{11}$$

依照上述的 N-gram 语言模型分词,然后比较出现概率,返回的是被认为最有可能的输入纠错结果。但是考虑到通常语料库越大,返回多个出现的概率相同的纠错关键词的概率就越大。所以对返回的出现概率相同的多个关键词再次分别建立 N-gram 模型,进行分词,每个返回的长度为 l 的关键词分布切分为 $3(l-1)$ 个子关键词,计算出每个子关键词在语言样本中的权重,最后相加得出总权重

$$W_i=\sum_{i=1}^{3(l-1)}tf_idf_i \tag{12}$$

式中,$1<j\leqslant n$,n 为返回的纠错关键词个数。

由上述公式分布计算并比较 n 个返回关键词的总权重,返回权重值最大的词,即为最终纠错系统的返回的正确关键词。

上述方法是对语言样本建立 N-gram 模型,并且比较了切分关键词出现的频率之后,再通过引入 TF/IDF 的方法,计算已经初步筛选出来的纠错关键词的分词后的总权重,最终比较得出最优化的结果。

3　实验结果及分析

上述提出的算法,实现了完全基于统计的方法对搜索引擎中的关键词输入错误进行检查和纠错。为了对在此提出的方法进行客观准确地评价,在不同的条件下模拟了检索中的关键词输入错误,设计了不同情况下适用的测试集。

在基于 N-gram 语言模型中,对于作为测试集的语言样本的选择很重要。根据不同大小的文件建立的语言模型,包含该语言的字、词的范围大小会有很大的区别。

根据纠错算法的规定,测试集选择了大量的词语作为搜索关键词,在语言样本测试集中做初始的查询,作为关键词测试集中的词语,是以随机的方式挑选,并且遵循以下的原则:

1. 按照两字词、三字词和四字词的仿真比例挑选;

2. 测试的关键词在挑选之前不考虑其正确或错误。

关键词测试集在不同大小的语言样本的测试集中的查询成功率,如图 1 所示。

对纠错系统的评价,以召回率(Recall,又称查全率)和准确率(Precision,又称查准率)作为评价标准,将其定义为:

Recall = 纠错系统返回的不为空的词的个数/关键词测试集中词的总数;

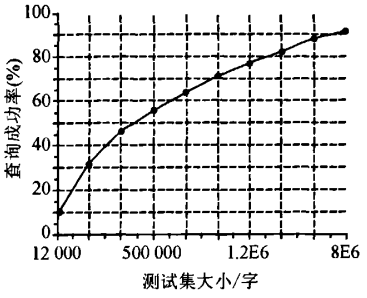

图 1 初始状态下的查询成功率

Precision = 系统正确查出的错误的词的个数/关键词测试集中词的总数。

将纠错系统应用于关键词测试集和语言样本测试集，通过第一步先建立 N-gram 统计模型，根据分词后出现频率比较，得到初步的结果（准确率 1，如图 2 所示）第二步，用 tfidf 方法计算关键词的权重，得出最后的统计信息结果（准确率 2，如图 2 所示）。

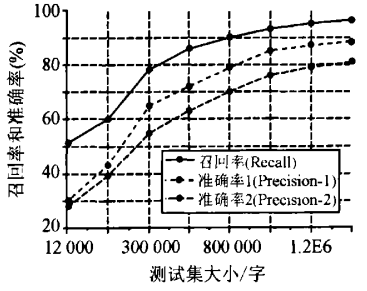

图 2 通过构建统计模型以及优化后的查询成功率

从实验结果的统计信息中可以得出以下结论：

1. 语言样本的测试集的大小对于纠错系统的准确率影响很大，测试集越大，得到的可以参考的统计信息越多，就可以提高对关键词纠错的准确率；

2. 通过进一步的分析关键词的统计信息，即计算切分后的词频和逆文档频率值，可以显著地提高纠错准确率。

4 结 语

随着搜索引擎在互联网中扮演越来越重要的角色，其技术也在不断地提高和完善，对于用户输入的关键词进行检查并且给出纠错建议，是 WEB2.0 时代下互联网技术向智能化发展的迈进。在此提出了一种完全基于统计信息的方法，将计算语言学中自然语言处理技术与中文的信息检索技术结合起来，对用户可能输入错误的关键词进行检查和纠错。基于统计的方法有着不可比拟的优势，因为其考虑到了人类的语言习惯，通过对语言样本上下文的统计信息分析从而得出更为准确的结论。同样的，自然语言也在不断的发展，本文提出的的算法还需要在考虑自然语言特点、对统计信息进一步的分析方面进行完善，以得到更加准确的分析结果。

参考文献：

[1] 胡晓清. 网络搜索引擎中文纠错功能实例剖析[J]. 图书情报工作，2008，1：1-6.

[2] BIGERT J. Probabilistic Detection of Context-Sensitive Spelling Errors[C]//Proceedings of LREC，2004：1-4.

[3] KUKICH K. Techniques for Automatically Correcting Words in Text[J]. ACM Computing Surveys，1992，24(4)：377-439.

[4] MAYS，ERIC，DAMERAI. Context-based Spelling Correction[J]. Information Processing and Management. 1991，27(5)：517-522.

[5] 施得胜，等. 基于统计的中文错字侦测法[J]. 电脑与通讯，1992，8：19-26.

[6] 张仰森，等. 基于二元接续关系检查的字词级自动查错方法[J]. 中文信息学报. 15(3)：36-43.

[7] 马金山，等. 利用三元模型及依存分析查找中文文本错误[J]. 情报学报，2004，6：723-728.

[8] 刘云，等. 现代汉语合成词结构数据库[J]. 现代化教育技术与对外汉语教学，2000，11：1-5.

作者简介

陈智鹏（1984－），男，河南卫辉人，硕士研究生，研究方向为信息检索、自然语言处理，E-mail：czpbupt@gmail.com；

吕玉琴（1945－），女，吉林长春人，教授，研究方向为通信软件智能终端；

刘华生（1984－），男，江西彭泽人，硕士研究生，研究方向为信息检索、自然语言处理；

刘 刚（1970－），男，天津人，博士，讲师，研究方向为网络安全、物理电子学；

屠 辉（1987－），女，宁夏固原人，硕士研究生，研究方向为信息检索、自然语言处理。