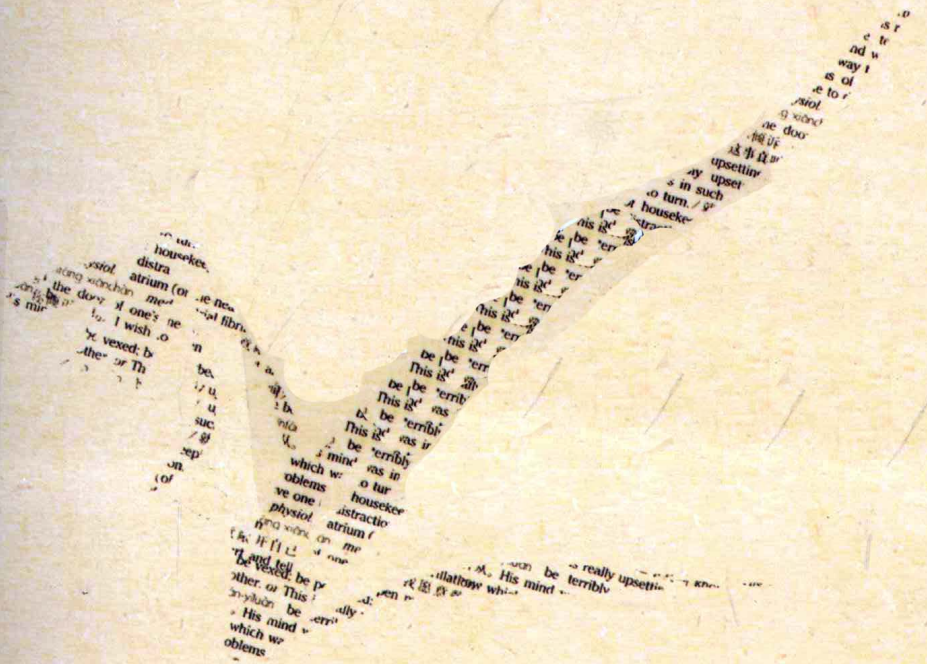
独秀外国语言文学博士文库

通过句法位置提取中文关键词的实验研究

王家钺⊙著



苏州大学出版社
Soochow University Press

独秀外国语言文学博士文库

通过句法位置提取中文关键词的实验研究

王家钺⊙著

苏州大学出版社
Soochow University Press

图书在版编目(CIP)数据

通过句法位置提取中文关键词的实验研究/王家钺著. —苏州:苏州大学出版社,2011.10
(独秀外国语言文学博士文库)
ISBN 978-7-81137-821-4

Ⅰ.①通… Ⅱ.①王… Ⅲ.①自然语言处理 Ⅳ.①TP391

中国版本图书馆 CIP 数据核字(2011)第 205306 号

书　　名:通过句法位置提取中文关键词的实验研究

作　　者:王家钺
责任编辑:谢永明
封面设计:大雅堂

出版发行:苏州大学出版社(Soochow University Press)
社　　址:苏州市十梓街 1 号　**邮编**:215006
印　　刷:宜兴市盛世文化印刷有限公司印装
网　　址:www.sudapress.com
E - mail:yanghua@suda.edu.cn
邮购热线:0512-67480030
销售热线:0512-65225020

开　　本:880mm×1230mm　1/32　**印张**:5.125　**字数**:180 千
版　　次:2011 年 10 月第 1 版
印　　次:2011 年 10 月第 1 次印刷
书　　号:ISBN 978-7-81137-821-4
定　　价:20.00 元

总序

广西师范大学外国语学院的“独秀外国语言文学博士文库”即将出版，这是广西外国语言文学界一个新的佳音，可喜可贺。博士文库是博士生在自己专业方面所写的文集，应当有较高的学术水平。所谓“博士”，除了特指一种高级学位（这正是“博士文库”中“博士”的含义）外，用更高的标准来看，还可以有另外一种解释，就是吴宓先生所说的“博雅之士”。我认为对现在的博士生的更为严格的要求，就是希望他们每人都能成为“博雅之士”。“博”就是知识广博深厚，有雄厚的专业知识和综合文化背景，“雅”就是为人高尚雅致，有崇高的人格。这样的知识分子就是合乎条件的博雅之士。广博雄厚的专业知识和综合文化背景当然重要，但它只是“博士”的必要条件，而非其充分条件。充分条件必须包括高尚的人格以及崇高的思想道德水平。博士论文则主要体现了博士生的学术水平。本文库是广西师范大学外国语学院博士的论文集，它是这个学院博士的思想结晶和劳动成果，应该受到学术和教育界的重视。博士生的学术研究需要出成果，本文库就是他们在学术上经过认真钻研思考后用自己心血创造的丰硕成果。知识分子应该具备在学术上创新和独立思考的能力，而检验和提高创新与独立思考能力的最佳途径就是写学术论

文。广西师范大学外国语学院编辑和出版本文库，说明了学院对培养浓厚学术氛围、提升学术水平的重视。事实上本文库的出版不但表现了学院对撰写学术论文的重视，而且也彰显了这些博士较高的学术水平。我很高兴看到本文库的出版，并且希望能有更多的高等学校出版博士研究生的论文集。说得更加富于诗情画意一些，我希望本文库能够成为给高校博士论文集出版带来第一个春天的第一只燕子。在这里，我满怀感谢之情向广西师范大学外国语学院和本文库的作者们致以深切的敬意。谢谢你们，加油，广西师范大学外国语学院的老师们和本文库的作者们！

贺祥麟

2011 年 9 月于南湖之滨

Preface

The majority of keywords in Information Retrieval are nominal expressions. The automatic identification and extraction of keywords from online texts amount to that of nominal expressions. Since nominal expressions are distributed in different syntactic positions, the natural question that follows is whether nominal expressions in different syntactic positions weigh differently in their contribution to the keywordhood, a question not yet well addressed before Dr. Wang Jiayue.

According to his research, the answer to this question is yes: nominal expressions not only in different syntactic positions, but also at different hierarchical levels bear different weight to the extraction of keywords. Moreover, there is also a discrepancy between the head nouns and nominals with modifiers, between nominal expressions at L1 and those at L2, between subject nominal expressions and the others, between nominal expressions in Ba-constructions and those in Bei-constructions.

Dr. Wang's experiments, findings and statistics may provide some new perspective for the improvement of the robust research and application of the information retrieval technology.

Ning Chunyan

前言

本书是基于我的博士学位论文 Chinese Keyword Extraction by Term Positions 改写而成。

我于1998—2001年在广东外语外贸大学读博，师从宁春岩教授。宁老师的主要研究领域是生成句法和语言习得，是一位坚定的唯理主义者，这也自然影响到了他在自然语言处理领域的立场，所以他一直对基于规则的自然语言处理非常感兴趣。我这篇论文的选题一部分来自他的想法，即在基于统计的自然语言处理方法日渐成为主流的背景下，探讨基于（句法）规则的方法是否对中文信息处理有帮助；其内容是考察句法位置因素对自动提取中文文本的关键词（keywords）是否有效。

本书详尽描述了自然语言处理尤其是信息检索的各种主要方法，并对信息检索领域“相关性”这一核心概念做了非常充分的回顾与文献综述（这部分内容单独发表在《现代外语》2001年第2期上），在此基础上提出了基于句法位置提取关键词的初步设想。通过小型问卷调查证明了基本名物性短语的重要性。在实验部分，作者以从某技术型网站搜集的小型文本库为实验对象，首先证明了不同文本位置以及不同句法层级上产生关键词的可能性有显著差异；随后使用向量空间模型（VSM）提取其关键词，然后对文本库中的基本名物性短语的句法位置进行手工标注，通过这些位置提取关键词，再将两者进行对比。实验表明，通过句法位置提取的关键词与通过VSM提取的关键词没有显著差异。最后作者对这些实验结果

进行了讨论。

需要说明的是：如同任何计算机科学领域一样，自然语言处理的发展完全可以用日新月异来形容。近年来新的理论、方法和技术层出不穷，而本书所描述的研究也存在许多不足之处，因此本书只能作为该领域的参考。另外本书省略了原附录中的计算机程序部分，并对各附录进行了重新编号。

香港城市大学潘海华博士和揭春雨博士、广东外语外贸大学桂诗春教授和曾用强教授等针对论文提出了许多宝贵意见。另外在我撰写论文期间得到了许多同学和朋友的热心帮助，包括广东外语外贸大学的陈鸿标博士、杨寿勋博士，黑龙江大学郭龙江，哈尔滨工业大学傅忠传等，在此一并致谢。文中的舛误概由本人负责，欢迎读者提出宝贵意见。

ABSTRACT

Keywords are the best content descriptors, more effective than other index terms for information retrieval (IR) systems, especially when the rapidly growing information sources are putting retrieval precision into highlight. Statistics-based IR and keyword extraction (KE) systems view documents as bags of unordered words, treating all index terms as equally important, without regard to their syntactic position. This paper tests the intuition that the syntactic position of Chinese nominal phrases is helpful for keyword extraction and compares the results with those of statistical KE.

Web pages can be treated much in the same way as normal text. Our investigation of some web search engines shows that their conceptions of relevance are different. Based on a detailed discussion of relevance, it is argued that there has not been a good link between the operability of system-oriented relevance and the rich achievements of user-oriented relevance studies. It is decided that topical relevance ought to be the attitude to be taken by web search engines and to be assumed in the present research. The approach to topic extraction based on human intuition is believed to be a promising direction worthy of efforts, because by extracting topic words, the subset of documents that really match the user's information need can be clearly determined, unlike the "standard" retrieval systems that only decide which documents are pos-

sibly relevant. Given that such human intuitions about relevancy can be well described, topically relevant results can be successfully retrieved and the outcome of the IR system will be more satisfactory.

We conducted a corpus-based study of (a) text positions—keywordhood and (b) syntactic positions—keywordhood relation. Attention is focused on Base NPs, which are manually annotated from a collection of technical documents, with their text positions (title, introduction/conclusion) and syntactic positions (subject, verb complement, etc.) marked according to a pre-designed scheme. The statistic results of the first experiment showed a high correlation between the Base NPs' syntactic position and their potential of being keywords. Subsequent experiments confirmed our expectation that text positions were helpful for KE, but syntactic position appeared not, which led to the conclusion that text position was more valuable than syntactic position with regard to KE.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Motivation

With the rapid growth of information in the modern world, especially with the aid of the Internet, the increasing number of available texts, mostly electronic, has highlighted the need to develop efficient IR techniques. The major concern of modern IR is to work on unstructured documents in natural language. Keyword extraction (KE) is generally believed to be an important technique that can increase IR systems' performance and is the focus of our present research.

IR techniques have been developing at an amazing speed, and some advanced models such as Vector Space and Probabilistic Retrieval (with many variations) have largely dominated the field. These new methods have shown surprisingly good performance and thus have been employed in various areas. The Vector Space Model, for instance, has been used by word-based web search engines such as Alta Vista. This is the reason why the VSM weighting will be used as an important method in our experiments. The majority of these systems are statistical in nature, originating from Luhn's well-known observation that, among others, "the frequency of word occurrences furnishes a useful

measurement of word significance." (Luhn, 1958) However, although these methods have been quite effective, human language is nevertheless rule-based instead of being subject to mathematical principles.

One very important fact is that full-text indexing, which is certainly the best method to facilitate string searching, is in many cases inadequate for the users, who "often use short topic phrases in exploring a collection" (Gutwin et al., 1998: 82) as can be seen in practices of digital library use and web search. Conventional IR seems to lose the field here to keyword-based retrieval.

As one of the most eminent application areas of language information processing, web search is quite different from experimental IR or KE. List-based search engines also provide a pre-compiled directory, but the huge number of web pages on the Internet is beyond the capacity of human review. The problem is comparable with that of KE: manual work always falls behind time.

The user can turn to different engines for different needs, but his patience is frequently worn out. What is more, few of the engines can satisfy one of the most frequent and simplest demands: "Find web pages about this or that," in the sense that the results returned to the user include too much irrelevant information. Most search results are links to web pages that contain the query expression. Moreover, the ranking of retrieved results is frequently confusing.

A question naturally arises: "What is relevance?" The question seems to have been answered from two major orientations. The system-oriented view is that relevance is equal to topicality, and that a document is relevant if its topic matches that of the query. Clearly, such an

understanding assumes that relevance is an objective notion, i. e. without the need to ask the user's judgments. The user-oriented opinion argues that relevance is a subjective notion, i. e. to be judged only by the user. Furthermore, such judgments are always dynamic and even indeterminate. This is certainly true. However, IR systems have to work with some justification. Many retrieval systems accordingly have a relevance feedback mechanism that allows the user to select "more relevant" results so that the system can fine-tune the searching process with minute modifications.

In the present research, a system-oriented position is adopted, with the assumption that there must be a steady understanding of relevance, and that the system can arrive at an objective document topic or group of keywords. This assumption works as our justification for regarding KE-based IR as a reasonable substitute for conventional approaches. Our focus, therefore, will be on the development of more effective methods of extracting keywords or topic words from documents. Relevance is here defined as the match between the query term and the keywords or topic words of the document.

Throughout this paper, "keywords" and "topic words" mean the same thing. That is to say we ignore the strict understanding such that "keywords" are predefined and "topic words" are dynamically extracted or generated from the documents.

1.2 Discussions on KE methods

Nominal phrases are the most significant index items for IR systems; they constitute the majority of the query terms used by search

engine users. A study of user queries on Excite (Jansen et al., 2000) shows that, of the most frequently used terms, most are nominal, although the study was not intended for this purpose. For this reason, these phrases deserve concentrated attention.

As we have mentioned, KE is usually done by statistical means (for details see Chapter 4). There are advantages in statistical methods. Firstly, "language (but not linguistic) analysis" can be done without much cost, because frequency calculations, presumably one of the simplest tasks for the computer, can be very easily implemented with any programming language. Secondly, theoretically speaking, with empirical data collected from the corpus, the "ultimate" language rules and regularities can be infinitely approximated although in actuality no one can confidently cry out the "eureka." Thirdly, since there has not yet been any systematic description of human language with real descriptive adequacy, statistical language study has gained more and more support. In fact, much of the work for statistical approaches is to deal with surface level phenomena rather than language internal rules. We are nevertheless interested in finding out if using linguistic analysis methods can obtain comparable performance or even better.

So, what possibilities are there for KE, apart from the statistical ones? Text positions, e.g. title, introduction/conclusion, have drawn much attention for KE. By increasing the weight of a keyword candidate according to its text position, the new weight will intuitively better reflect the candidate's importance in the text than does the original weight. This assumption has actually been widely adopted in keyword extraction and yielded satisfactory results.

Our intuition tells us that not all NPs in a sentence make equal