自然语言语义理解下的信息检索模型

吴晨[1,2]，张全[2]，缪建明[1,2]，韦向峰[2]

（1，中国科学院研究生院，北京 100039；2，中国科学院声学研究所 北京 100080）

摘 要： 为了提升信息检索系统性能，语义学被引入了传统基于概率统计算法的检索模型之中，本文将沿着这一思路，以HNC自然语言理解体系为基础，阐述HNC理论下，引入语义知识，逐步构建检索系统的策略。并结合已经取得的部分研究成果对这一策略进行佐证。企望能为基于语义和理解的检索系统的发展作引玉之砖。

关键词： 信息检索，自然语言处理，语义，概念层次网络

Information Retrieval based on the Content Understanding

Wu Chen[1,2], Zhang Quan[2], Miao Jianming[1,2], Wei Xiangfeng[2]

(1Graduate School of the Chinese Academy of Sciences, Beijing, 100039; 2Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100080, China)

Abstract: In order to improve the performance of the present IR (Information Retrieval) systems. Semantics was introduced into the statistical IR models. This paper will discuss a development schema, which, based on the HNC (Hierarchical Network Concept) natural language understanding theory, will help to implement such a semantics-based IR system. Some research results will be brought out in the paper. These results are able to well illuminate and demonstrate the feasibility and effectiveness of the schema. We hoped that the proposed method would do benefit to the development of the new monobasic retrieval systems.

Keywords: Information Retrieval、Nature Language Processing, Semantics, HNC

1 引言

基于概率统计算法的信息检索策略在过去的十年中取得了巨大的成功[1-5]。具有里程碑意义的搜索引擎也因此诞生了。然而，随着应用的深入，人们开始发现当前的检索技术存在着不尽人意的地方，检索结果往往求全不求准，文章中与检索词相同的文字符号的数量直接影响了检索结果。究其原因，这主要是由于目前检索系统抛弃词语之间存在的语义关系，忽视词语在句子表达中所起的作用引起的。自然而然我们会想到，能否让语义信息有效的为检索系统服务，而不是将解题思路局限于把词语看作是孤立个体的数学算法上。如果能够做到这点，必将会从根本上提升信息检索的性能。本文将就这一问题，以HNC自然语言理解体系为基础，结合其在信息检索方面新近取得的研究成果，从实现策略的角度，讨论在自然语言理解技术支撑下的信息检索模型。

本文第二节介绍HNC中基于形式化语义、服务于信息检索的自然语言理解技术；第三节在已取得成果的基

基金资助：973项目“自然语言理解的交互引擎研究”（2004CB318104）、中科院声学所知识创新工程项目“HNC语言知识处理理论及技术”

作者简介：吴晨（1979～），男，北京，博士生，研究方向：自然语言处理；Email:wuchen@mail.ioa.ac.cn

础上介绍第二节理解技术支撑下的检索模型及实现方法；第四节给出小结。

2 HNC 中的自然语言理解模型

自然语言理解实质上是发现语言文字符号所表达的“义”的过程。只有让计算机把握了自然语言所表达的内涵，我们才有可能实现高性能的信息检索。

如何让计算机看懂这一内涵，这就涉及到了语义表示的问题。HNC 设计了语义基元网络来提供形式化、概念化的“义”所必需的参照系。HNC 用概念树作为语义网络的基本组成单元，用概念树上的概念节点作为语义描述的基本单位，用树与树以及树内上下节点之间的关系来对概念之间的关联性进行描述。图 1 为 HNC 概念树的一个例子，该分支为其所管辖的语义范畴提供了规范和参照，具体定义了专业活动（professional activities）下的部分概念。

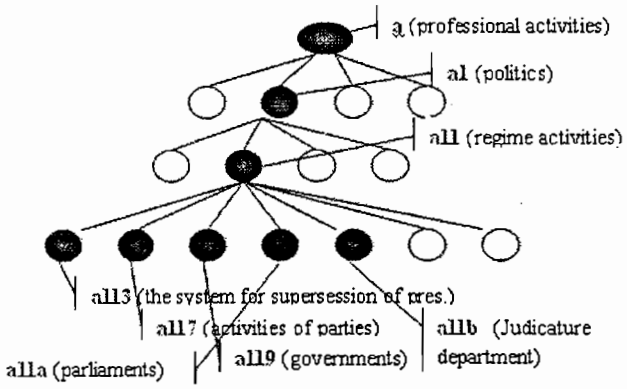

图 1 专业活动概念树的一个分支

提供了概念化的语义内容表示方法后；第二步就是实现自然语言符号向这一表示的映射（Mapping），即自然语言的理解。为了说明问题，我们将本文介绍的信息检索模型与语义 Web（Semantic Web）[6]在思路、方法及策略上进行对比。应该说，语义 Web 与本文所倡导的模型具有相同的解决问题的思路。两者都希望借助语义手段，通过赋予计算机可被其识别的文本的语义解释，来最终提高计算机对于文本的理解处理能力。作为实现语义 Web 的重要组成部分的语义标注（Semantic Annotation）[7]完成的就是用本体（Ontology）对“义”进行解释和标注的任务。然而，本文所研究的信息检索的实现策略和出发点与语义 Web 存在差异，语义 Web 希望创建者在创作网页时就根据某种标准为内容提供语义标注，进而能被计算机利用，而本文主张的信息检索更多是在已有信息资源的基础上的自动语义标注，进而服务于检索。这一在已有资源上做检索的思想与目前的信息检索的做法是一致的。这一思想可以在继承目前信息检索已取得成果的基础上增强信息检索的效能。可见，为了实现本文主张的信息检索，我们必须对已有资源进行自动加工，自动语义标注（Automatic Semantic Annotation）的工作必不可少，而自然语言理解所要完成的重要内容就是自动标注，标注的依据为上下文所蕴含的语义。根据语言中语义承载单位以及理解难易的不同，我们将这一理解过程划分为了两个阶段：句子理解和句群理解。

句子理解服务于词语的理解，因为在句子的具体语言环境中，词语的意义才能得到具体的体现。词语理解可以看作是将词语所蕴含的内容向语义网所定义的概念基元进行映射的过程。沿着这一思路，句子理解也必然需要一个承载句子含义的形式化的符号体系来作为理解的最终表示形式。HNC 定义了这一符号体系—句类（Sentence Category）[8]。句子理解的目标是用有限的句类表示式来表示句子的语义结构，同时获取构成句子的各个语言单位的语义。为此，HNC 定义了 57 组基元化的基本句类表示式，以及 57*56 个混合句类表示式[8,9]来表示句子的内涵。

图 2 给出了一个经过理解处理后所得到的句子理解结果，处理结果包括了句子的句类表示式（SCE）以及各元素所映射的概念（Term concept）。HNC 用于这一处理的技术称为**句类分析技术**。

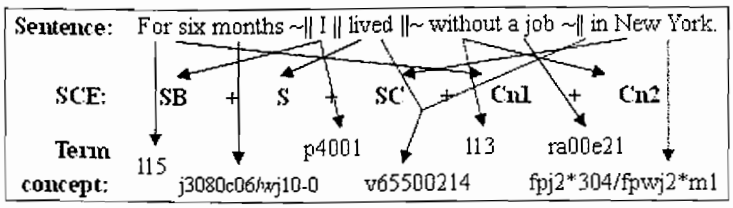

图 2 HNC 句子级理解示例

句群理解是在对句子进行理解的基础上进行的，也就是依靠句子理解输出结果来进一步处理句群。

与句子理解的处理方法一致，HNC 同样通过构造能够承载句群含义的形式化的符号体系来作为理解输出的框架。这一框架被称为“语境单元”。语境单元由 3 部分组成：领域（DOM）、情景（SIT）和背景（BAC）。领域描述了句群语义中心的类型，它将对应到语义网络定义的某一个概念上，共定义有 108 个主领域[10]。情景则是对表示内容的展开说明，它由领域句类表示式来表示，领域句类表示式是对句群语义结构的一个形式化表示，这与句类表示式对应句子是相同的。背景则是对句群中涉及的时间、地点等信息的概念表示。

图 3 给出了一个句群理解的示例，图中输入为两个自然语言语句，输出为经过理解处理得到的句群理解结果，理解结果用被赋予了知识的语境单元框架表示，图中以 SGU（Sentence Group Unit）标记语境单元框架的部分。HNC 用于这一处理的技术称为**领域单元萃取技术**。

3 语义理解下的信息检索模型

第 2 节中提到，本文主张的信息检索与语义 Web 在出发点和策略上有所不同，我们的目标是在现有信息资源条件下实现高性能的语义检索，而语义 Web 主张的更多的是在建立内容的同时添加语义标记，从而使得计算机具有理解内容的能力。语义 Web 并不注重对文章的理解，而注重对本体标准的定义以及利用本体对内容进行标注的研究。所以它没有提到如何在目前条件下解决基于语义的信息检索的问题。本文所要讨论的信息检索将会与其有所差异，它是一个在现有资源条件下，在句子和句群理解的基础上实现的检索。一方面它涉及被检索文本的理解问题，另一方面涉及检索条件与理解结果之间的匹配问题。

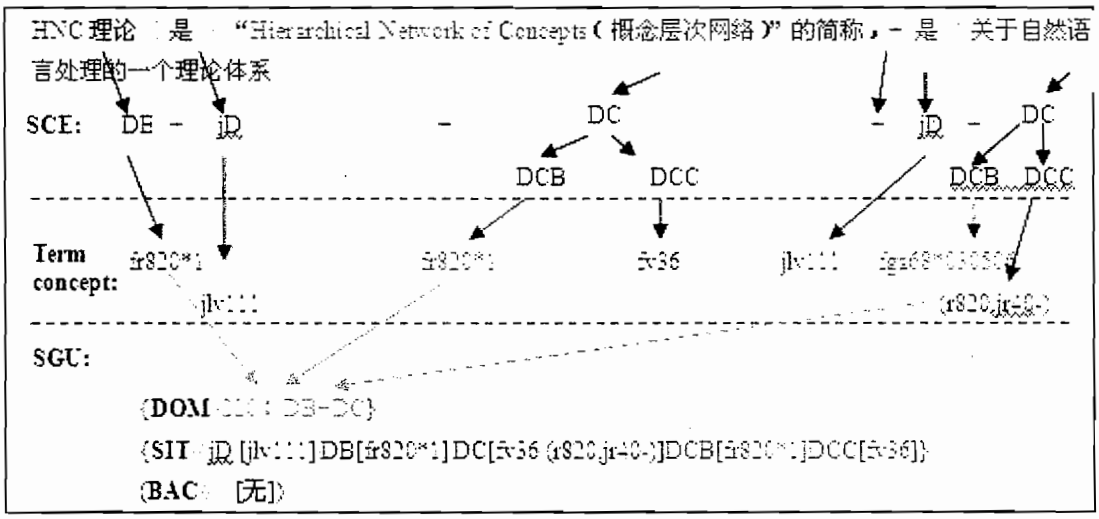

图 3 HNC 句群理解示例

根据目前 HNC 所具有的文本理解能力，我们将着重讨论已经初见成效的基于句子理解和句群理解的信息检索模型。

3.1 语义与概率统计方法相结合的信息检索模型

基于语义的信息检索与语义 Web 一样，不是一蹴而就的工作，所以以 HNC 自然语言理解处理为基础，我们

提出了分阶段实现的策略，根据句子理解、句群理解的特点，将其与传统统计模型结合来实现检索。实现方案如图 4 所示。

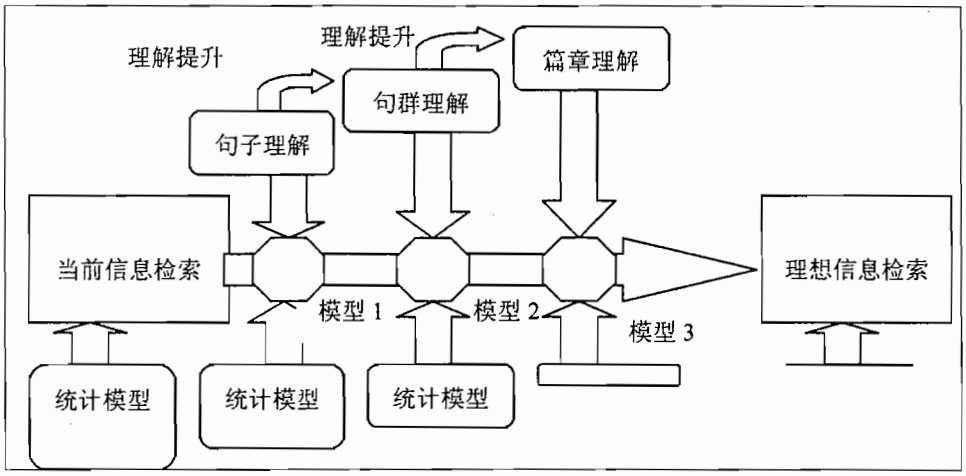

图 4　息检索实现策略

图中箭头标示的是信息检索发展的一个趋势，箭头下方为统计模型在整个信息检索模型中所占的比重，可以看到，沿着箭头所指的发展方向，统计模型所占比重在不断减少。箭头上方为计算机对文本的理解度，沿着箭头所指的发展方向，理解的层次越来越高。我们在箭头上标示了三个点，分别对应基于不同文本理解程度的 3 个模型。本节将介绍已经取得一定研究成果的模型 1 和模型 2。

3.1.1 模型 1

仅仅依靠句子理解，还无法脱离开统计模型，它能够帮助系统快速实现从句子理解到篇章理解的跳跃。引入句子理解后，信息检索模型将具有一些新的能力，包括：依据句子理解的结果获取词语语义并用概念加以标示，根据词语在句子中的功用赋予词语不同的权重，可以利用概念之间的相关性和语义网络概念树的特点对文本进行有指导的分类，提高检索准确率。

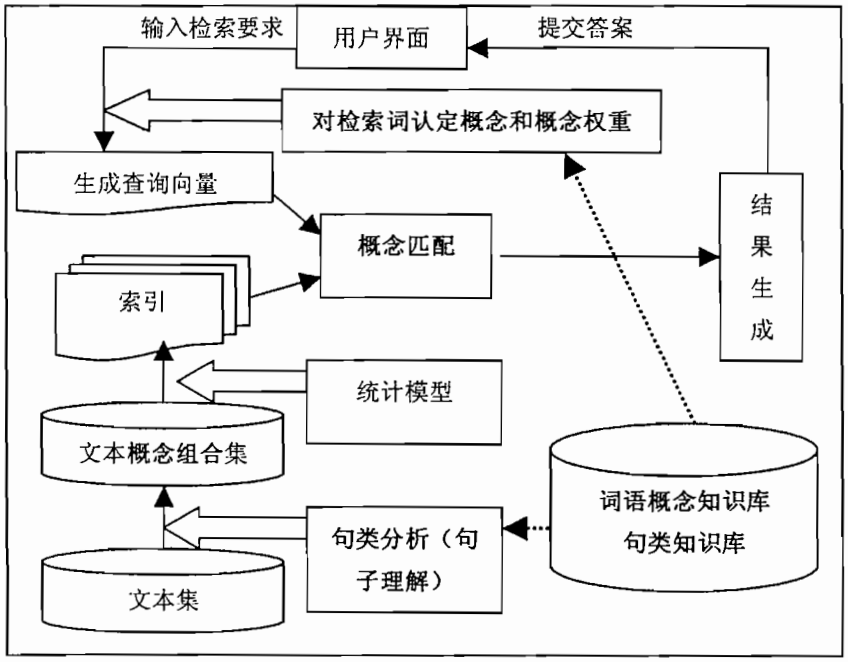

图 5 基于句子理解的信息检索模型框图

该模型的基本构建思路在于：以句子理解为基础，通过句类分析系统实现准确地词语切分，同时获取词语概念，再根据词语在句子中的位置赋予相应概念不同的权重，最后用统计模型对概念进行处理。图 5 给出了一个该模型的框图。从图中可以看出，索引的过程中加入了句类分析（句子理解）的部分，该部分替换了传统中文统计模型中的分词。实质上，句类分析兼具分词和获取词语语义的功能，句类分析后的文本集将形成一个以篇章为单

位的词语概念组合，概念是通过句子理解来获取的。统计模型则负责对概念建立索引。可采用的统计模型比较多，包括 TF-IDF[11]、K-means 聚类[12]以及语言模型[5]。采用这些统计方法的概念模型都初步证实了比同等条件下基于词语的方法效果要好，尤其是基于聚类的方法[13]。该方法以语义网络所定义的具有强领域信息的概念树中的若干概念构成分类的种子，通过 K-means 聚类算法，依靠种子，实现有指导的文本聚类，避免传统 K-means 算法中无指导迭代的盲目性。模型最后再根据聚类分布，实现检索，取得不错效果。

与传统的统计模型相比，检索请求和索引之间的相似度计算（Similarity Calculation）被新模型中的概念匹配所取代，这主要是由于在新模型中使用了概念作为处理的中介，这同时也要求系统必须能够将用户输入的检索请求用概念的方式来表述，这一方法通过发掘用户输入关键词间的语义关系可以很好的帮助系统明确用户意图，并且在无语义歧义的情况下对概念进行扩展，最终形成概念查询向量。

3.1.2 模型 2

基于句群理解的检索模型也保留了统计模型，作为从句群到篇章的重要理解过渡手段。该模型的特点在于：通过句类分析准确切分词语，获取词语的语义；根据句群分析的结果得到句群所属的领域，如政治、军事、法律等等共 108 大类。根据领域信息对文章进行分类，基于分类实现检索。该模型同时也具备信息过滤的功能，依据是句群分析结果中的句群立场信息。

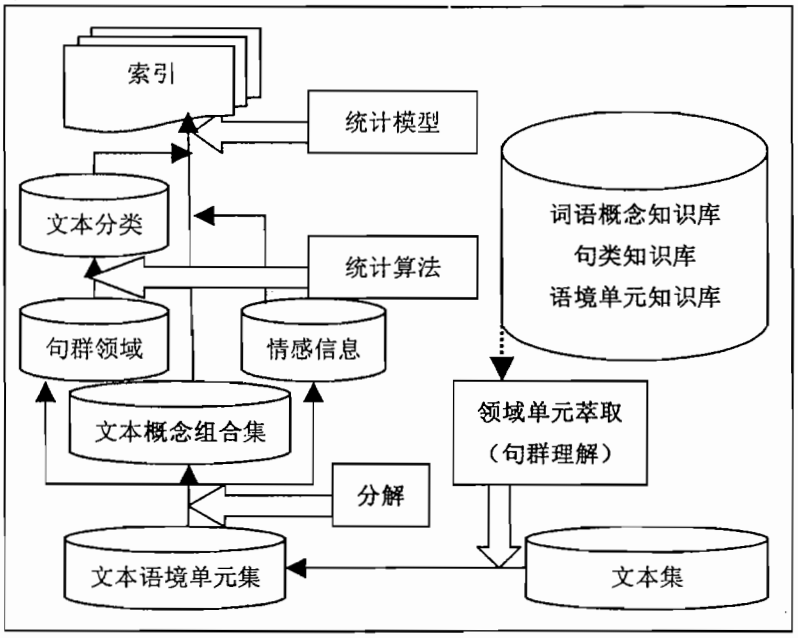

图 6　基于句子理解的信息检索模型框图

构建该模型的基本思路是以句群理解为基础，根据句群理解得到的语境单元框架中的领域信息，对构成句群的文章进行分类，给出属于每类的置信度，同时根据每类出现每一概念的可能性计算出文章出现每一概念的可能性。最后根据语境单元框架中包含的关键概念运用统计模型再次对概念索引。最后用两个索引进行插值平滑出最终的主索引，具体索引算法可参见[14]。图 6 给出了该模型索引部分的框图，其余部分与图 9 相似。从图中可以看出，在对文章进行索引的过程中加入了语境单元萃取（句子理解）部分的内容，这样就使得我们在得到句群的领域信息的同时获得情感信息，情感信息只在文本过滤时使用，它表征了作者在写句子时表达的情感，比如支持、反对某个事物。具体算法可详见[15-16]。通常情况下，信息检索只用到领域信息。

4　小结

从模拟人脑理解的角度出发，实现一套完全以模拟人脑语言智能为主体的检索系统决不是一蹴而就的事情，然而，这又是解决目前信息检索存在问题的重要解题思路，本文从这一现状出发，以 HNC 自然语言理解体系为基础，结合其已经取得的在理论和工程上的研究成果，给出一种语义理解与统计方法相结合的信息检索发展策略，希望能为智能信息检索的发展提供一些有益的信息。

参考文献：

[1] Bailey, P., Craswell, N., & Hawking, D. Engineering a multi-purpose test collection for Web retrieval experiments [A]. Information Processing and Management, 39, 853–871. 2003.

[2] Lalmas, M. Logical models in information retrieval: Introduction and overview [A]. Information Processing and Management, 34(1), 19–33. 1998.

[3] Miyamoto, S. Fuzzy sets in information retrieval and clustering analysis[M]. Kluwer Academic Press. 1990.

[4] Salton, G., & McGill, M. J. Introduction to modern information retrieval [M]. New York: McGraw-Hill. 1983.

[5] C. Zhai and J. Lafferty. Two-stage language models for information retrieval [A]. In Proceeding of SIGIR 2002.

[6] Bemers- Lee T. Semantic Web road map[EB/OL]. http://www. w3. org//design issures/semantic. html, 1998.

[7] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval [A]. Web Semantics: Science, Services and Agents on the World Wide Web, Volume 2, Issue 1, 1 December 2004, Pages 49-79. 2004.

[8] Huang Zengyang. HNC (Hierarchical Network Concept) Theory [M]. Beijing: Tsinghua University Press. 1998.（In Chinese）

[9] Miao Chuanjiang. Guide of HNC (Hierarchical Network Concept) Theory [M]. Beijing: Tsinghua University Press. 2005. (In Chinese)

[10] Huang Zengyang. Mathematics and physics symbol system of language in language concept space [M]. Beijing: Ocean Press. 2004. (In Chinese)

[11] Salton, G., & Buckley, C. Term-weighting approaches in automatic text retrieval [A]. Information Processing and Management, 24(5), 513–523. 1988.

[12] J MacQueen. Some methods for classification and analysis of multivariate observation [A]. In: Proc of the 5th Berkeley SympMath Statist and Prob 1. California: University of California Press ,281-297. 1967

[13] Wu Chen, Zhang Quan. An Information Retrieval Method Based on Language Concept Space Using Clustering Method [A]. Computer engineering. 2006. (In Chinese with English abstract)

[14] Wu Chen, Zhang Quan. Content matching: a concept-based approach for information retrieval [A]. Journal of southeast university (English edition), 12(5). 2006.

[15] Jing Yaohong, Miao Chuanjiang. An Algorithm of Extracting Text Character Based on a Model of Context Framework [A]. Journal of computer research and development, 41(4). 2004. (In Chinese with English abstract)

[16] Jing Yaohong. An information retrieval method based on language concept space using clustering method [A]. Computer engineering and applications, 39(13). 2003