语义理解下的自然语言处理及信息检索模型*)

吴 晨[1,2] 张 全[2] 缪建明[1,2] 韦向峰[2]

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

摘 要 本文从如何利用语义来构建理想信息检索系统的角度出发,以HNC自然语言理解体系为基础,给出了一种在这一体系下分步构建信息检索系统的实施方案。结合HNC在信息检索方面新近取得的研究成果,从理论和工程两个角度介绍了这一体系服务于信息检索的相关内容,讨论了其中几个重要模型的实现方法。最后沿着方案思路给出了在目前成果基础上进一步发展检索系统的展望。

关键词 信息检索,自然语言理解,语义,HNC

Nature Language Processing and Information Retrieval Based on the Content Understanding

WU Chen[1,2] ZHANG Quan[2] MIAO Jian-ming[1,2] WEI Xiang-feng[2]

(Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)[1]

(Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China)[2]

Abstract Aiming to construct an ideal information retrieval (IR) model based on semantic methods, this paper reports on a development and implementation schema. This scheme is based on a nature language understanding theory (Hierarchical Network Concept theory, HNC in short), and also points out the main procedures in constructing such an ideal IR model stage by stage. On the one hand, this paper introduces the related HNC theory to illuminate the schema. On the other hand, it provides some research results to demonstrate the feasibility and effectiveness of the schema. At the same time, some expectations have been addressed.

Keywords Information retrieval, Natrare language understanding, Semantic method, HNC

1 引言

信息检索的发展到现在已经经历了两代。第一代信息检索是人工分目录检索;第二代是计算机依靠算法实现的以关键词检索为主要表现的自动信息检索。第二代检索的主要技术特征为概率统计算法在检索中取得的成功[1-9]。这一技术的出现无疑奠定了检索技术发展的一个重要里程碑。采用概率统计算法实现检索的最大优势在于:已有相当多成熟、可供利用的数学模型,可以为其提供扎实的理论支持,同时算法实现的方案明确、复杂度不高、易于工程化。然而从另一角度看,这种方式也不可避免地存在处理上的欠缺,因为数学模型本身是抽象、无物理含义的,当它用于某个具有特定含义的应用模型时,必须符合"应用题"成立的前提条件。然而,在信息检索中,以篇章为单元、以词语为单位的数据集合必然不能满足目前所采用的各种概率统计模型对数据无关性的要求。同时,词语之间存在的语义关系以及词语在句子中所起的作用又是我们可以利用的丰富处理资源。目前的信息检索技术对应用前提的回避及对语义内容的漠视已经成为影响其应用效能进一步提高的瓶颈。因此,信息检索技术不能局限于把纯数学算法作为核心的思路,应该把重点放在如何合理而有效地利用语义内容上,这是从根本上提升信息检索性能的关键。本文将从HNC自然语言理解体系出发,结合其在信息检索方面新近取得的研究成果,从语义的角度,讨论在自然语言理解技术支撑下的信息检索模型,着重讨论基于局部语义理解(语义理解和统计模型相结合的处理方法)的信息检索策略,并对基于全文理解的信息检索进行展望。

本文第2节介绍HNC中形式化的语义表示方法-概念;第3节介绍基于形式化语义、服务于信息检索的自然语言理解技术;第4节在已取得成果的基础上介绍第3节理解技术支撑下的检索模型及实现方法;最后给出小结。

2 HNC的语义表示方法

HNC(概念层次网络)理论是中科院声学所黄曾阳先生创立的自然语言理解体系。体系可以划分为两部分处理内容:一是概念化的语义内容表示;二是自然语言符号向这一表示的映射(Mapping),即自然语言的理解。本节将介绍第一部分内容,第二部分内容将在下一节讨论。

语义是蕴含在语言文字符号下的隐现内容,语义表示方法的最终目的则是为内容提供形式化的表示手段,也就是提供将隐现的"义"转化为显现的"义"的载体。HNC中用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描述概念之间的关联性。HNC共定义了456棵概念树,概念以树中节点的形式存在,每棵树中的概念都具有相同的特性,父节点涵盖的概念范畴比其子节点广。

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。节点注释中的字母即为该节点的概念标识,括弧里的内容为语义解释,如a表示专业活动,a1表示政治(politics)。这些都是抽象的概念,并不代表具体实例。当要表示具体实例时,需将一个或多个抽象概念相结合。如"美国政府"表示为"a119+fpj2*304","a119"在图1中已经标示。fpj2*304为特指的美国。图2为对应图1

*)本文承国家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"的资助。吴 晨 博士,研究方向为自然语言处理;张 全 研究员,博导;缪建明 博士;韦向峰 博士,助理研究员。

中概念树的工程实现，它构成了 HNC 概念基元知识库。

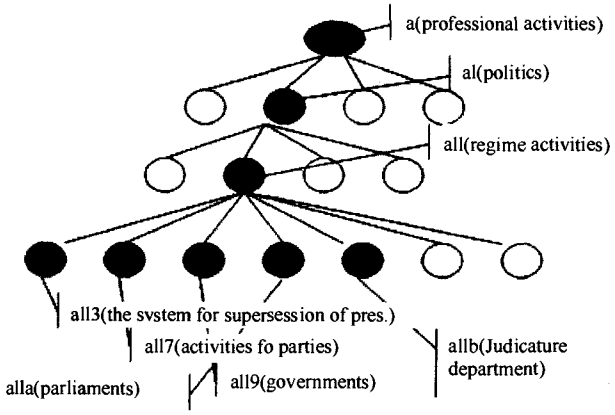

图 1　专业活动概念树的一个分支

可以看出，以 HNC 语义概念基元网为基础的知识表示工作主要在于对领域知识的抽象表示，即发现以概念树为领域单元的抽象概念、概念的继承层次、潜在的关系和公理等。对于领域中的实例，在建模时并不予以考虑。这与本体(Ontology)构建的思想相同。实际上，HNC 定义的概念就是一种通用本体(Generic Ontology)，它涉及自然语言表示的语义内核，涵盖多个领域，同时又将表示本体(Representational Ontolog)和任务本体(Task Ontolog)[10]融入其中，可以描述事物的实体、事件涉及的动态知识等。HNC 概念建模所最求的目标就是抽取和概括事物的共同特点，并且以特定符号串的形式加以标识。

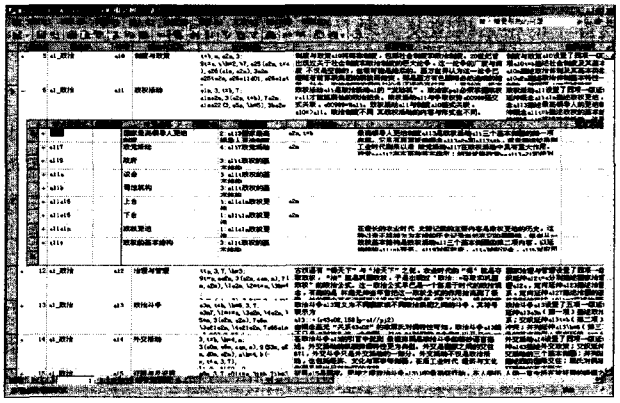

图 2　专业活动概念树实例

3　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。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)[11]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)[12]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。要完成一个自动标注的过程，一方面需要依靠词语本身自带的信息，另一方面则要依靠词语所在的上下文信息。这就引出了本节“句子理解”和“句群理解”的内容。同时，句子和句群是自然语言中最为重要的两个语言单位。抓住这两个语言单位的语义核心，必然会给信息检索带来巨大的帮助。

3.1　句子理解

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)[13]。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[13,14]来表示句子的内涵。

HNC 进行句子理解的技术被称为句类分析技术[13,15]，这一技术的数学模型可归纳为目标公式(1)。

$$SCE=f(SCH_1,SCH_2,\cdots,SCH_n)$$
$$SCH=g(TC_1,TC_2,\cdots,TC_n) \quad (1)$$

式(1)中，SCE 为句类表示式，SCH 为语义块(对应于句类表示式中预定的槽，类似于短语，由词组构成)，TC 为词语描述对象的概念。$f()$ 和 $g()$ 为变量之间的约束函数，其物理意义为词语所表述对象的抽象概念以及与其所处句子间的约束关系。

句子级的理解可以归纳为以下步骤，如图 3 所示。

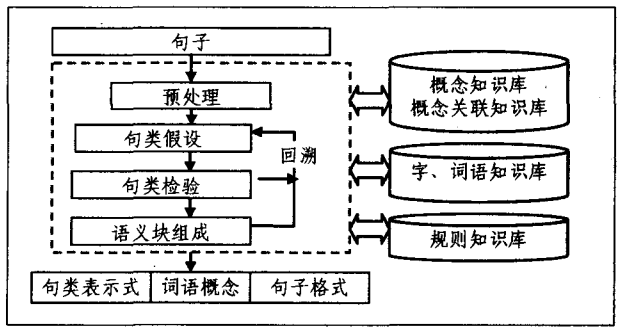

图 3　HNC 自然语言理解句子级处理模型

图 3 中虚线框内的部分为处理环节，输入为自然语言的语句，输出为句类表示式以及词语所描述对象对应的概念。处理的第一个环节为预处理，如中文处理中的分词。第二个环节为句类假设，该环节完成的功能为根据句子中“可疑”词语蕴含的概念，对可能会产生的句类进行假设。第三个环节为根据其他词语所蕴含的概念知识，对句类进行断定。第四个环节为根据第三个环节作出的判断，以语义块为单位对词语进行断定。如果第三或者第四个环节通不过，那么处理会

引入一个回溯的过程，重新进行“假设”和“检验”，直到成功。处理的输出包括了“句类表示式”、“词语概念”以及“句子格式”，句子格式类似于构成句子的主谓位置规定，差异往往由“被”字句、“把”字句等不同表示习惯引起。

句子理解需要多个知识库做支持，为“假设-检验”提供支持，这些知识库包括“字、词语知识库”、“概念关联知识库”以及“规则知识库”。图 4 给出了词语知识库的一个工程实例。

图 4 中包含的“义项数”为词语可能对应的概念数；“Hnc 符号”项为词语所对应的概念；“句类代码”项为该词语所能激活的句子的句类表示式代码。它们是假设检验的基础。知识库的构建文在文献[15]中有详细的说明。

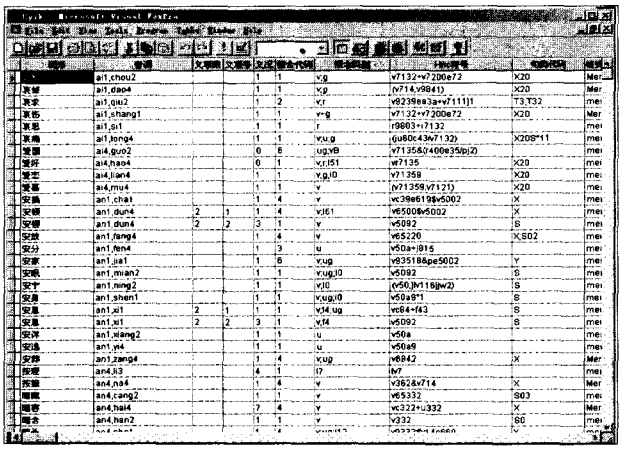

图 4 词语知识库示例

图 5 给出了一个经过理解处理后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)。

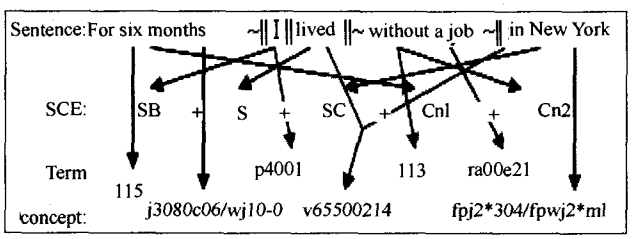

图 5 HNC 句子级理解示例

3.2 句群理解

自然语言理解的最终目标是以一种计算机可理解的方式对篇章语义进行描述。然而，从句子理解向篇章理解的跨度太大，所以引入了句群理解的层级。句群是一组由某一语义中心统摄的、关系密切的句子的集合体。从语言的表达看，如果表达的意思比较复杂，往往不是用一个句子，而是用几个句子组成的句群来表达，分成几句话来说。这样要比用一个结构很复杂的句子更容易让人理解，这也是形成句群的主要原因。抓住句群语义表达的中心，必然会为理解篇章带来帮助。

HNC 中，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[16]。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。

句群语义理解的步骤可以归纳如图 6。

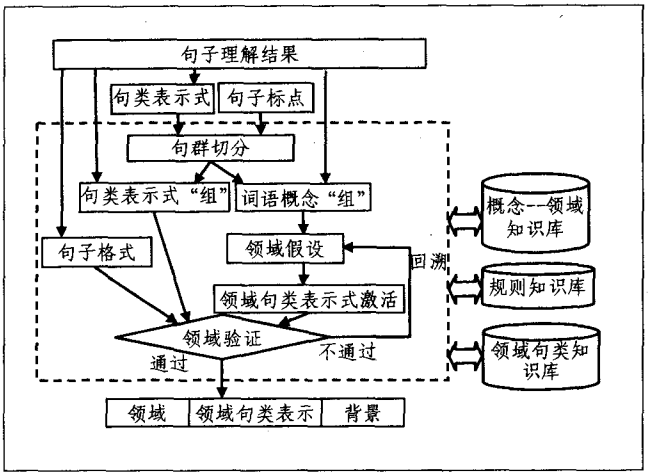

图 6 HNC 自然语言理解句群级处理模型

图 6 中虚线框内的部分为处理环节，输入为句子理解的结果。首先根据句类表示式以及标点符号对句子进行句群划分[17]，再以划分的句群为单位，根据句群中概念的领域信息强弱，假定句群所属领域，同时根据领域提取领域句类知识库中对应的领域句类表示式，最后将表示式中的内容与组成该句群的各个句子的句类表示式与格式进行一致性认定，最终确定领域以及领域句类代码，同时根据句类表示中的背景信息叠加成句群的背景信息。

图 7 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。

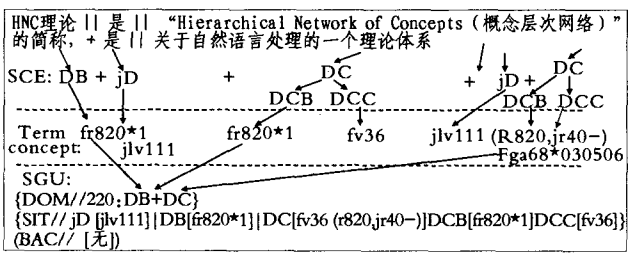

图 7 HNC 句群理解示例

4 基于理解的信息检索技术

第 3 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。

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，同时对未来基于篇章理解的检索模型进行思考和展望。

4.1 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 8 所示。

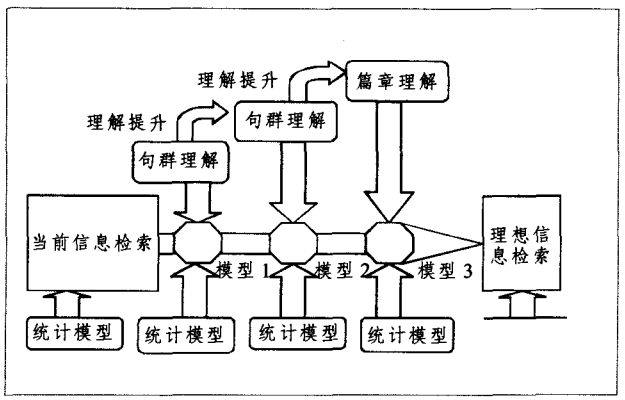

图 8 信息检索实现策略

图 8 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，模型 3 将在下节中给出思考和展望。

4.1.1 基于句子理解的检索模型

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，包括可以正确地切分词语，同时依据句子理解的结果获取词语语义并用概念加以标示，并可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。

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的词语切分，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 9 给出了该模型的一个框图。从图 9 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[18]、K-means 聚类[19]以及语言模型[8,9]。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[20]，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。图 9 中的词语概念知识库即为图 4 所示的库。

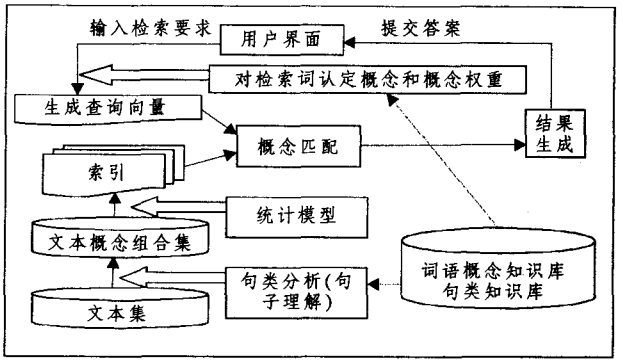

图 9 基于句子理解的信息检索模型框图

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，文献[21]中介绍了这一方法。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地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。

4.1.2 基于句群理解的检索模型

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。

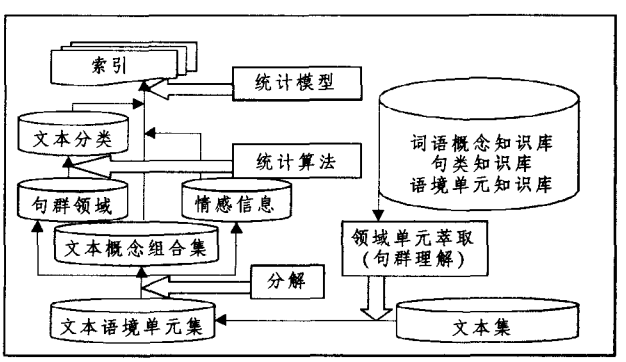

图 10 基于句子理解的信息检索模型框图

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。然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文献[22]。图 10 给出了该模型索引部分的框图，其余部分与图 9 相似。从图 10 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。具体算法可详见文献[23,24]。通常情况下,信息检索只用到领域信息。

4.2 基于篇章语义的信息检索模型思考

理想的信息检索模型作者认为必须满足最基本的三点要求:第一,对查询请求的处理不局限在关键字方式上。信息检索中,只用关键字来表达用户的查询是不够精确的。因此,查询处理必须具有理解用户叙述的能力。在理解叙述的同时,有针对性地扩展查询项,同时增加约束条件,包括位置约束、语法约束、种类约束等等。所谓位置约束反映的是所描述对象及立场在文章中出现的前和后约束条件。语法约束表示检索对象是否该句主语,是否是施事者等信息。种类约束表示是地名还是人名,还是其他等。有了这些约束及其表示,信息检索完成约束条件的检测,从而提高查询的查准率。第二,具有对文本内容进行抽取和表示的能力。文章必然有它描述的主要内容、体现的中心思想和背景立场,这些信息埋藏在语言文字符号之下,要准确地进行信息检索,必然需要对这些信息进行发掘和归纳。同时,系统还必须具有对所抽取内容进行形式化表示的能力,必须满足无歧义和完整性的要求,最终能够服务与检索。第三,具有点面结合的匹配策略。用孤立的词条来完成对文本集的索引是不够周全的,词条间的关联性被切断。同时,词条索引是一维的,词条与文本的联系无法被几何地体现,这种条件下实现的检索只能完成关键词的匹配。在理想信息检索中,应该能够充分考虑关键词限定的点和语言环境铺设的面之间的关系,进而考虑约束条件的问题。所以,在文本内容抽取和表示的同时,系统就需要既对文本内语义凸现的点进行标注,又注重对文本内容面的理顺,形成一个针对文本的具有层级的语义描述框架。在匹配时,则点面结合,在保证重心相符的同时,也保证语义吻合,实现多层级的内容匹配,在内容发生抵触时,给出提示。

HNC中,基于篇章语义的信息检索将有可能成为一种理想的检索方式,检索模型的设计充分考虑了以上所介绍的三个基本点。模型由三部分内容构成:用户需求的理解及概念化表示、文本内容的萃取、概念联想脉络出发的文本匹配。模型框图如图11所示。从图中可以看出,统计模型已经从模型中淡化,语境生成被纳入了模型中。语境生成的作用就是4.1.2节中统计模型所作的工作,负责实现完全意义上的从句群到篇章理解的飞跃,也就是实现在语境单元萃取(句群理解)基础上的文本内容萃取,并最终生成针对文本内容的概念化摘要。本文着重介绍HNC文本内容的形式化表示方式,它将是构成HNC信息检索模型的核心。这一概念化的表示框架被称为语境框架[16],由五部分组成:对象与内容的描述(BCN//BCD)、作用与效应的描述(XYN//XYD)、过程与转移的描述(PT)、关系与状态的描述(RS)以及背景(BAC)。

BCN//BCD是对语境单元框架中DOM的一个综合;沿用DOM的分类方式,是对文本所属领域的界定依据。XYN//XYD、PT、RS则是分别对文本描述内容中产生的作用(施动情况)与效应(产生影响)、过程(事件发展的线条)与转移(局部的效应)、关系(产生的对象之间的联系)与状态(产生的中间态)进行描述。这些信息由语境单元框架中的SIT归并、迭代得来;BAC为文章描述的背景信息,它与语境单元框架中的BAC相对应。

语境框架是概念化表示的,它一方面提供了描述文本内容的基本依据,另一方面也构成了文本索引的基本单元。

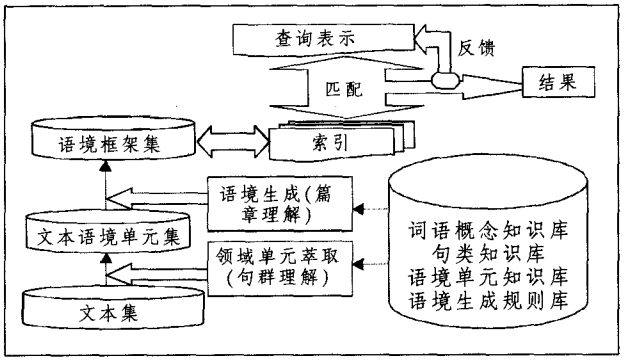

图11 基于篇章语义的信息检索模型

那么,如何从语言单元框架集演绎得到语境框架?这将构成本文下一步要研究的内容。

结束语 信息检索系统的出现为人们高效地获取信息提供了可能,成为了现代信息技术史上的一个重要里程碑。然而,伴随着人们对检索要求的不断提升,我们不得不去寻找更为理想的下一代信息检索实现方法,建立在自然语言理解技术之上的检索自然成为人们关注的焦点。然而,若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。所以,本文从这一现状出发,以HNC自然语言理解体系为基础,结合其已经取得的理论及工程研究成果,给出了一种信息检索系统的发展思路,提出了以HNC体系为根基的理想信息检索一种实现策略,企望为这一技术的发展作引玉之砖。

参考文献

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

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

[3] Meadow C T, Boyce B R, Kraft D H. Text information retrieval systems. 2nd ed. San Diego, CA: Academic Press,1999

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

[5] Crestani F, Pasi G. Soft computing in information retrieval. Germany: Physica Verlag and Co,2000:102-121

[6] Salton G. Automatic information organization and retrieval. New York: McGraw-Hill,1968

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

[8] Zhai C, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval//Proceedings of SIGIR. 2001:334-342.

[9] Zhai C, Lafferty J. Two-stage language models for information retrieval//Proceeding of SIGIR,2002

[10] Gamper J, Nejdl W, Wolpers M. Combining Ontologies and Terminologies in Information Systems//5th International Congress on Terminology and Knowledge Engineering, Austria, 1999

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

[12] Kiryakov A, Popov B, Terziev I, et al. Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web, 2004,2(1):49-79

[13] Huang Zengyang. HNC (Hierarchical Network Concept) Theory. Beijing: Tsinghua University Press,1998(in Chinese)

[14] Miao Chuanjiang. Guide of HNC (Hierarchical Network Concept) Theory. Beijing: Tsinghua University Press,2005

[15] Wei Xiangfeng. The software platform for expanded sentence category analysis based on the HNC theory. Doctor's academic dissertation of IOA, CAS. Available: http://www. hncnlp. com/Abs/absEwxf. htm (in Chinese with English abstract)

[16] Huang Zengyang. Mathematics and physics symbol system of language in language concept space. Beijing: Ocean Press,2004 (in Chinese with English abstract)

[17] Wu Chen, Zhang Quan. Some rules for detecting Chinese sentence groups in nature language processing. Computer engineer-

ing, 2006 (in Chinese with English abstract)
[18] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 1988 24 (5): 513-523
[19] MacQueen J. Some methods for classification and analysis of multivariate observation//Proc. of the 5th Berkeley SympMath Statist and Prob 1. California: University of California Press, 1967:281-297
[20] Wu Chen, Zhang Quan. An Information Retrieval Method Based on Language Concept Space Using Clustering Method. Computer Engineering, 2006 (in Chinese with English abstract)
[21] Wu Chen, Zhang Quan. Introduce Linguistics and Semantics into the Probabilistic Indexing Method for Chinese IR: Using Concept Theory. Proceedings of ICTAI-2006
[22] Wu Chen, Zhang Quan. Content matching: a concept-based approach for information retrieval. Journal of Southeast university (English edition), 2006, 12(5)
[23] Jing Yaohong, Miao Chuanjiang. An Algorithm of Extracting Text Character Based on a Model of Context Framework. Journal of Computer Research and Development, 2004, 41(4)
[24] Jing Yaohong. An information retrieval method based on language concept space using clustering method. Computer Engineering and Applications, 2003, 39(13)

(上接第 72 页)

点发送的监测信息包大小为 256 bits。

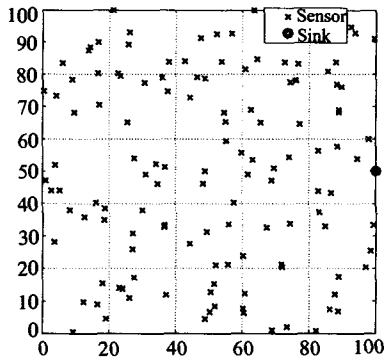

图 3 仿真环境俯视图

设定每一个节点的初始能量相同，均为 3J，并采用自由空间传输信道模型进行仿真[3]。在该模型中，节点发送和接受数据时器件的能耗 $E_{elec}=20nJ/bit(1nJ=10^{-9}J)$，发送功率放大器的能耗为 $E_{amp}=100pJ/bit/m^2(1pJ=10^{-12}J)$。鉴于以上设定，传感器节点发送消息时消耗的能量 $E_{tran}=E_{elec}+E_{amp}=20nJ/bit+100pJ/bit/m^2$；节点接收数据时消耗的能量 $E_{recv}=E_{elec}=20nJ/bit$。

网络平均能耗仿真结果如图 4 所示。其中，图 4(a)和图 4(b)分别为运行定向扩散路由和运行蚁群节能路由的传感器网络在 10 个小时内的能量消耗平面图。图 4(a)中网络的平均能量消耗为 214.3mJ，图 4(b)中网络的平均能量消耗为 143.75mJ，减少了 32.9%。相对定向扩散路由，蚁群路由节约能耗的主要原因是：定向扩散路由是一种基于查询方式的路由协议，在路由建立阶段要进行兴趣消息的泛洪，这样增加了建立路由的代价；而蚁群节能路由通过节点的分布式计算来有选择地发送路由建立消息（在无线传感器网络中，节点的计算能量代价远远小于通信能量代价），不但减少了通信数据量，而且在建立路由时就把“节能”作为选取标准，因此路由能耗得以大幅降低。与此同时，由于蚁群节能路由在建立过程中综合考虑了节点能量及节点间通信能耗，网络中的能量负载也更加均衡了，这可以从图 4(a)与图 4(b)的对比中反映出来。

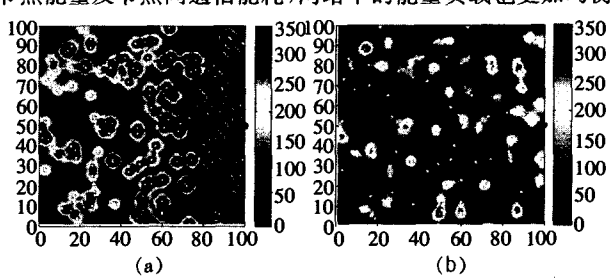

图 4 路由协议能量仿真结果平面图

3.3 关于网络生存时间的仿真

在网络生命周期仿真中，仿真环境与上小节设定相同，网络持续运行直到有节点能量耗尽。基于定向扩散路由的网络共计运行 57 小时 22 分，基于蚁群节能路由的网络共计运行 84 小时 04 分，这意味着基于遗传蚁群路由的网络生命周期比基于定向扩散路由的网络延长了 46.5%。基于蚁群节能路由的网络生命周期得到大幅延长的原因有二，一是分布式路由优化计算使得节点节约了大量的通信能耗；另外，蚁群节能路由在建立过程中考虑了能量负载在网内的均衡性，避免了由此局部节点能量耗尽而引起的网络出现盲点，这也是网络生命周期延长的主要原因。

结束语 本章中针对无线传感器网络的特性，设计了一种蚁群节能路由协议。该协议具有分布式计算、分布式信息存储、路由通信代价少及本地存储信息量小、低能耗、能量负荷均匀等特点。另外，该路由协议还具备适应性好，可扩展等特点。

为了具有较好的通用性，在上面所述的路由协议为平面路由，即无线传感器节点之间在路由上没有层次之分。如果针对其他应用环境，该协议可以方便地扩充为层次路由协议，即网络可以根据位置信息化分成为动态变化的若干簇，簇内通过能量状况及所处位置选举产生一个簇头节点。簇内的簇头节点在其任期内保持侦听状态，其他节点在无数据时保持休眠，当有数据时才唤醒并将数据直接发送给簇头，各个簇头节点再通过蚁群路由形成二级最优通路与 Sink 节点通信。通过这种简单的扩充蚁群节能路由就可以在不同的应用环境中获得更加优越的性能。

参 考 文 献

[1] Marco D, Vittorio M, Alberto C. The Ant System: Optimization by a Colony of Cooperating Agents. IEEE Trans. on Systems, Man. and Cybernetics-Part B, 1996, 26(1): 1-13
[2] Rappaport T S. Wireless communications, principles and practice. Prentice Hall, 1996
[3] Friis H T. A note on a simple transmission formula//Proc. IRE. 1946,34
[4] Intanagonwiwat C, Govindan R, Estrin D. Directed diffusion: A scalable and robust comm unication paradigm for sensor network//Proc. 6th Annual Int'l Conf. on Mobile Computing and Networks (MobiCOM 2000). Boston, MA, August 2000