中文检索与汉语语义概念图表示*

陆汝占

上海交通大学 计算机科学与工程系 上海 200240

E-mail: rzlu@sjtu.edu.cn

摘 要：当今信息时代，人们从海量信息中获取所需要信息已成为日常生活的组成。人们普遍感到缺憾的是检索准确率低，这将限制手机检索的应用前景。问题的症结在于检索系统采用布尔模型“与”、“或”运算这类“离散型”方法处理语言，分裂割断了词语在概念上的联系和完整性，从而造成噪声。对此，如何从用户需求及网页摘录的表达式重构语义概念关系构成完整整体是一个富有希望的构想。与当前结构主义语法和概率统计方法为主流强势所不同的，特别关心汉语是义符文字，直接显现内涵语义，词语构造直接对应概念耦合的特点，这些特点与印欧语所采用的 Ontology、Semantic Web 方法不同。本文探讨用汉语内涵概念图表示方法对汉语表达式重构概念图以确保概念关联的完整性。

关键词：中文检索，汉语语义概念图

Chinese Information Retrieval and Chinese Conceptual Graphs Representation

Lu Ruzhan

Department of Compute Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240

E-mail: rzlu@sjtu.edu.cn

Abstract: In the information age, it is a daily need for people to retrieve needed information from mass data. A general problem is the low precision of retrieval, which will limit the application of mobile search. The crux of the problem is that current IR systems are 'discrete' models such as a Boolean model manipulating operators like 'AND' or 'OR', which results in noise because of the splitting of the relation and integrity of concepts. Consequently, it is a promising idea to reconstruct the semantic conceptual relations from the expressions of requirements and snippets to form integral units. Different from current mainstream of structuralism grammar and statistical methods, this paper focused particularly on the characteristics of Chinese language. The characters of Chinese are ideographs, which display intensional semantics directly. The construction of words corresponds to the coupling of concepts directly. These characteristics show difference from the Ontology and Semantic Web methods applied in Indo-European languages. This paper studied how to reconstruct conceptual graphs from Chinese expressions using the representation of Chinese intensional conceptual graph, which would ensure the integrity of conceptual relations.

Keywords: Chinese information retrieval, Chinese Conceptual Graphs

1 引言

自然语言模型是一种非单调逻辑的类典型原型理论[1]。这是就语言结构与指称实体之间的对应关系而言的，实体类中除包括典型实体之外，还有非典型的实体。对于语言结构与语义概念（包括内涵的语义特征）之间的对应关系、理据来说，汉语有一个明显特点，就是：汉语是义符

*基金项目：国家自然科学基金面上项目（NO. 60873135）

作者简介：陆汝占（1940-），男，上海交通大学计算机科学系教授、博士生导师. 主要研究方向：汉语语料库加工技术、汉语内涵逻辑模型及其应用、基于概念内涵的智能检索、对话理解系统、语义 web。

文字，不同于印欧语系的音符文字，词语的结构与意义以名词为中心，不同于印欧语系的动词为中心，词语构造方式是毗连、组合，对应了概念的直接耦合。因此，与音符文字的语系相比，汉语的优势在于：内涵概念显露于结构外形，即字面上就可解读出来。要问这样的特点及其研究对于当前信息领域有什么实用价值呢？研究方法和成果是属于语言、哲学单纯文科领域还是属于文、理、计算机与信息的交叉领域？是否有实际应用背景赋予这项研究以生命力呢？

本文有感于林毅夫先生所谈的真正的创新起点不在于方法在于问题和现象，认为中文检索首先要考虑汉语的现象及其特点，而且特别关心汉语的语义概念及其分析，以及其上的操作运算。因此，本文在对郭绍虞先生的汉语文法和修辞[2]，以及王力先生的关于汉语合成词结构语义解读和“解字组词”[3]等语言学理论研究基础上，针对中文信息检索中存在的问题，探讨如何用汉语内涵概念图表示方法对汉语表达式重构概念图以确保概念关联的完整性，并对汉语语义概念图在中文信息检索应用中涉及的一些相关问题进行研究。

2 中文信息检索的现状、问题及解决思路

网页检索准确率现状与预期

以下面几个需求为例：

(a) “上海哪儿能买到火车票”

(b) “中国大陆新发现油田”（类似的有“发现新油田”）

(c) “高科技孵化基地”

(d) “红鸡蛋”

搜索引擎处理需求表达式中的关键词，作为主要概念依据，但不难发现返回的结果“不要的太多，要的太少”，准确率低下，夹杂了相当数量的用户不需要的信息、噪声。上述四例中，a 的检索结果会出现有人从“上海”签证到印度在孟买“火车站”，购“火车票”。而 b 的结果则出现“中国政府与尼日利亚合作在尼日利亚开发油田”、1963 年发现大庆“油田”、“中国大陆新发现”化石等。c 的需求中“高科技”是借代指称专业为高科技化的企业、单位，“孵化”是比喻意指扶植培养，培养对象是前者企业、单位，“基地”指称单位或地区，使用高科技手段孵化禽类是另一种需求的解读。d 的“红鸡蛋”有红壳鸡蛋、喜蛋，红的蛋黄，苏丹红污染的蛋这三种不同指称，“红”所赋予的概念属性不同，关联的隶属属性名称不同，会出现诸类同显关键词但实质内容不相关的检索结果。

此外，上述信息检索中出现的准确率不高的现状，也难于适应未来手机检索所要求的高准确率的挑战。所以期望能在搜索返回与用户之间加上一个再分类，将搜索引擎返回结果区分分类成：准确解/近似解/噪声。如是，则有望进一步提高现有检索的准确率，也有可能将前几条准确解返回手机检索用户，以适应手机检索中面临的高准确率的要求。

◆ 问题所在和原因

基于语言独立性假设，技术层面上是检索采用布尔模型，将词语中的关键词分割，再用“与”、“或”布尔操作（三个关键词至少有七种不同分布）。语言层面上用户原本用关键词、问句、完整语句好端端地表达了一个完整的需求，表达了自身的意愿，结果被分割成支离破碎的概念片，甚至孤零零的字（词）！完全割裂肢解了用户的意愿。这样的处理对于计算机来说是最快、最现实的，但处理结果并不一定理想。

◆ 解决思路

提高检索准确率的关键在于探讨用哪些办法来克服准确率低下的问题。眼下有希望的、也许是唯一能指望的是在语言层面上考虑如何准确地完整地表达用户与网页之间相应概念及其匹配，以补救技术上的缺憾。

1. 针对关键词（组）被“分割”，反其道行之设法构造“耦合”的整体｛<E，A，V>｝。

用户需求中关注的是有关某个实体（人、事、物）以及相关的必要特征，组合起来构成一个整体。用实体（类名 E），及其特征：属性名 A，其值 V，构成序偶组｛<E，A，V>｝。可以标引成树形或网形，联成整体称为概念图。事实上，网页信息中同名不同特征的、同特征不同名的现象很多，不完整的特征更是造成噪声的根本原因。

2. 针对计算机是单纯字符串匹配运算，设法在中文字符上赋予汉语语义概念及其关系的信息标引，将单纯字符串匹配加载为概念及其关系上的匹配运算。实际上是将词语表达式中的文字，按“解字组词”思想解读出内在的概念含义，字符之间的结构毗连转换、提升为概念之间关系的联结，列出关联理据，这才保证上述整体是特定的，能从海量信息中区分出特定实体，以满足用户的真实需求。于是计算机单纯字符串匹配运算转化为概念图匹配，最终是序偶组之间匹配。

3 概念检索的主要步骤

概念检索的主要步骤包括：用户需求的概念分析和概念图标引，网页检索摘录的概念分析及其概念图标引，最后是两个概念图的匹配、比对。

3.1 用户需求标引

用户需求，如关键词表达（稍复杂点的是用简短疑问句查询所需要的有关实体的信息，但再复杂的是语段陈述方式，这涉及语段文本分析。），提取概念及其关联的关系。通常用户所需查询什么样的实体信息，实体名及其特征值（如“红苹果”）会表达清楚，但特征的属性名不显现表达，如“黄香蕉苹果”，所指称的“苹果”（皮）颜色黄，味道香蕉味，“皮”、“颜色”、“味道”这类属性名，通常在日常生活中已熟知默认的，不言自明的，但是计算机检索中歧义和差错往往出在此处，例如“孵化”究竟是孵禽蛋还是比喻扶植培育意思，出现差异造成噪声导致检索准确率的低下。正确分析用户需求中的概念及其关联关系，然后标引成树、网形或图，称为概念图1[4-7]。图 1 是需求实例“[上海[哪儿]]能[买到[火车票]]？”及其对应的概念图。

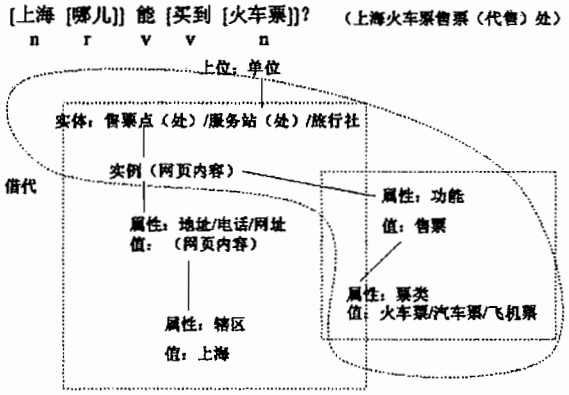

图 1 “上海哪儿能买到火车票？”需求对应的概念图

该需求求解的实体单位名称不具体、未知，可能是“火车票（铁路车票）代售（预售）

处（点）”，宾馆及单位的票务中心。具体地址、电话、网址未知，但所求单位的地址的约束条件是“地域：上海”，功能（职能）是“预售（代售、销售）火车票（铁路车票）”。需求关注：单位名称，地址信息。借代是指在需求问句中以所求解实体单位地址（疑问词“哪儿”）来指称该单位，修辞方法。另一个隐含的借代是以具体功能（售火车票）来指称单位。具体求解时要还原，既求单位名称，又求地址信息。

3.2 检索摘录的标引

检索摘录的标引，一般比用户需求标引复杂和困难，因此，可以在用户需求中所关注的概念关键词的导引下来分析和标引。换句话说，需求中所出现的关键词及概念表达<E,A,V>，也指望能在摘录中找到。如是，则优先比对匹配；如果有所短缺，那就会在相似近似解的范围内了。正确分析概念及其关联关系，生成概念图2。

3.3 概念图匹配

概念图1、2匹配。为使算法有效、简便，可以先将概念图1、2中的序列组{<E,A,V>}摊平，于是树、网形、图即转换为序列组了。序列组1、2的匹配类似于生物基因比对，可沿用生物计算方法[8, 9]。

4 概念图标引

4.1 概念图标引工程化的不同阶段

一个新方法的形成，大致经过：构想、原型、形成方法、最终出软件。初始阶段的构想和可行性思考至关重要。概念图标引最终作为工程来完成，至少前面要经历以下历程：起始、后继、最终目标走向。

①起始：手工标引+辅助软件、工具+可读电子词典（语言、专业两大类）。

人力手工标引的起始阶段是不可省的。开发有效的辅助软件工具，能辅助解读词典的释义项，根据释义中的语法结构、用词搭配的规律，生成若干模板、模式，用以抽取释义项中的概念，因为通常的语言词典并不是“概念词典”，释义时需要手工从词语表达式中抽象出概念，从经验上说，这“抽取”、“抽象”两个步骤是有难度的，但又十分重要。即使下一步从语料中统计学习，也还缺少不了这二者的干预和调整。

词典释义项可经过分词和初步句法结构分析后[6, 10-12]作概念分析[13, 14]。

②后继：建立词汇标引库，主要是单、双音节实体词如“笔”“车”，以及双音节复合词。

大部分工作可以借助于词典的释义，设法用辅助工具自动抽取必要的概念及其关联关系，构建<E，A，V>概念图[7, 15]。单、双音节实体词是原子结构（或称“基元”），由此再组成的复杂的复合名词（如“奥运金镶玉笔”）可看成是由“笔”组成的“金笔”，以及更大更复杂的颗粒结构。基于汉语是义符文字，组合方式毗连对应了概念直接耦合的特点，复合词所指称的实体及其概念内涵特征（属性名、属性值）其实是在中心词所对应的原子结构的概念内涵特征上作关于属性名、值的复合计算。如果将中心词（原子）的概念内涵特征作为实体类的共性的话，那么复杂复合词的概念内涵特征会有继承类共性特征以及凸显的个性特征两部分组成。其中个性特征，区分于共性特征的地方，有对某共性特征的变异，属性名不变，仅值更改：增加、删除、改变（变异）；另外也有增加新的属性名及值。如“笔”类中奥运金镶玉笔，功能：“写字作画”之外再增加“珍藏”功能，此外，组成、部分及材料上变异为汉白玉材料作杆，黄金作笔头、笔尖。但它还属于笔类，类的主特征未变异、未突变。这子类“奥运金镶玉笔”还属“笔”类中的典型子类。

复合名词的概念标引所基于的运算，类同于数理逻辑中的变量代入赋值、变量改名、删除、

增加。更复杂的计算在于修辞理据的计算，有些类同于范畴代数中的归约化简步骤。由此可见，词汇库标引以及复合计算方法作为后继目标是有望实现的。

③目标走向：努力朝自动标引、用户需求与摘录自动匹配方向迈进。我们已经在概念检索模型、概念关系构建、需求的自动标引等方面做了一些有益的探讨[4-7]，初步说明了我们设想的可行性。

目前并不限于用规则方法还是统计方法来实现[4, 5, 7]。重要的是从现象和问题出发，看看有什么办法和思路来解决所遇到的障碍。这是创新的第一步起点。未来的工作是进一步形成方法，逐步走向工程化、软件开发。

4.2 实体命名、概念及其关系标引方法的依据

实体命名本质上是编码，汉语以最小、最少的编码最贴切地指称最大、最广泛的实体对象类。举例，《牛津高阶英汉双解词典》插图 A1 中有关 bread 面包类，其中九种面包：牛角面包、小圆面包、一条面包（枕头面包）、切片面包、面包皮、面包片、圆面包，都是“面包”加上形状特征：条、片、圆或者比喻牛角；或部件组成特征：皮、切片；或内容材料特征：填馅、奶酪番茄的填馅面包，奶酪番茄面包。

而英语命名是完全不同的音符字母串，由此不难想象为什么要用 ontology 及 semantic web[16] 这样的网络将它们硬性归为面包（bread）类，犹如一只箩筐框起来，除此别无他法。汉语命名一般有规律：种类＝类＋种差，用到语言表达上就是：实体子类名＝类名＋凸显必要区分特征。

汉语命名特点给我们一个有益的启示，研究创新的起点及秉持的特色应该从汉语本体理论特点出发，而不是从普适世界各语言文化的共性出发。

4.3 概念图标引内容

概念图的表示框架为<E，A，V>，即对应的实体、属性名和属性值。

汉语构词及概念内涵特征表征是有规律的，结构上：复合词=基本词汇（原子）+语用复合成分；概念上：实体子类名（子类概念特征）=类名（类概念特征）+凸显必要区分特征；

例如：笔、金笔、奥运金镶玉笔（其概念图分别见图 2、3 和 4）。其中“金笔”实际上是实体类“笔”的一子类，该实体子类的名称是由实体类名“笔”加上“金”凸显了必要的区分于“笔”类中其他子类（如“毛笔”、“铅笔”）的特征，“金”是“金笔”的部件——笔头、笔尖这两者的材料（黄金合金，或铱金合金）。“金笔”继承了其上位实体类“笔”的特征——功能用途：书写/绘画。“奥运金镶玉笔”，在继承其上位实体类“金笔”的特征时，变异的是“组成”：笔头、笔尖——材料：“黄金”，还有“部件——笔杆——材料——和田玉”，以及功能用途：书写/绘画/珍藏，最后增加“珍藏”是价值所在。

“笔”的指称类可分为三个子类，其中两类具体实体，一是钢笔、毛笔、铅笔等典型类，另一类是“电”笔（试电笔、测电笔）等，非典型类，这类取名为“笔”是用修辞手段比喻而成，以笔的外形（包括杆、尖、尾等部位的形状）象典型类中的笔，其功能、用途已从典型类笔的书写作画变异为测电了。再有一类是文笔、败笔、伏笔、命笔、妙笔等，用修辞手段中的借代方法来命名的，其中名称所指称的是书法、绘画、诗文等作品或作品中的部分内容，或者作者写作行为，具有“妙”、“败”等好坏评价结论，无论是作品及内容以及写作行为都是用“笔”这个工具的，故看做多个递归借代过程：笔(E1)→功能(A1)→写字/绘画(V1, E2)→结果(A2)→作品(V2, E3) / 作品局部(V3, E3)。最后一个实体 E3，作为评价的对象，其结果为成功或不成功，即妙、败等属性值。三个子类可见图 5，分别为典型类、非典型类，以及相关概念关系。

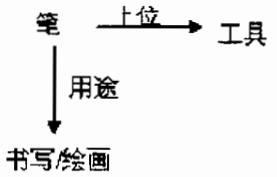

图2 “笔”概念图

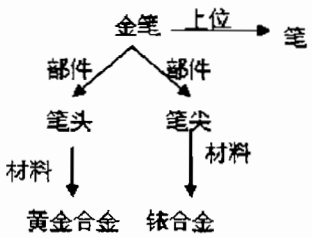

图3 “金笔”概念图

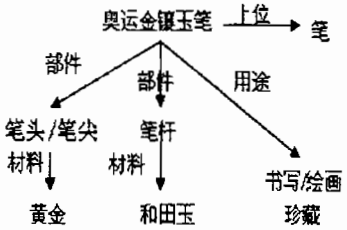

图4 “奥运金镶玉笔”概念图

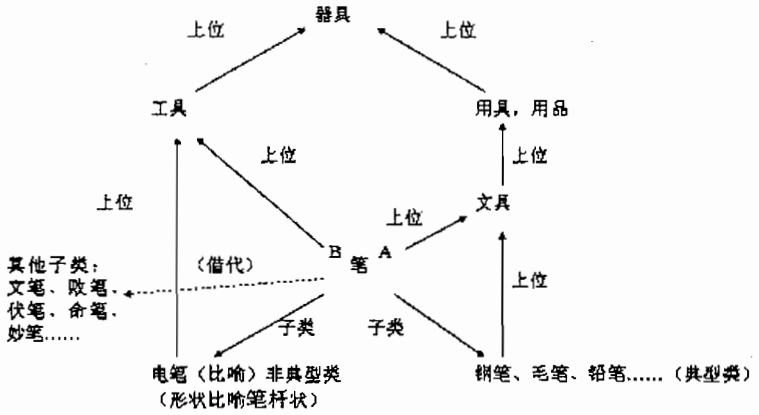

图5 “笔”指称分类概念图

5 未来工作

语料中出现的实体名E，属性名A，属性值V，三者的表达式不一定完全、完备，时有缺省。这根本上会影响实体概念及其关联的完整性，从而影响检索准确率。此外，也是基于统计方法的用户评价分析、情感倾向分析研究中致命的瓶颈。<E, A, V>缺省情况有三种：<E?, A, V>；<E, A?, V>；<E, A, V? >。分别缺省E、A、V。作为概念图标引方法来说，重要工作就是要解决如何求解、补偿缺省成分。

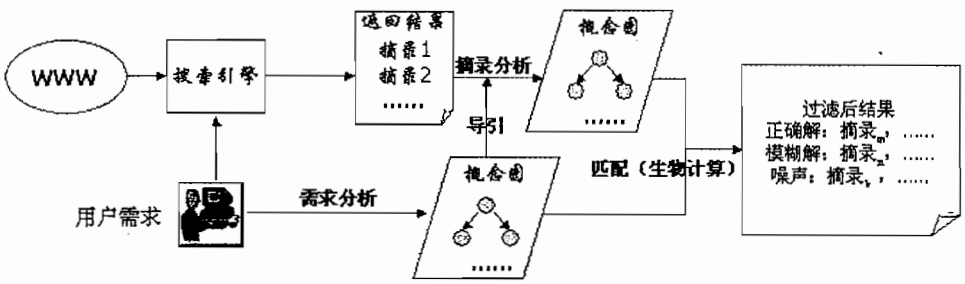

图6 需求概念图导引下的信息二次检索流程图

针对布尔模型在信息检索中出现的问题，试图将关键词所对应的相关概念连成概念图，以实现概念的整体完整性，并用于信息的二次检索以期提高检索的准确率。基于概念图的信息二次检

索基本思路是：如图 6 所示，尝试在搜索引擎的外层增加一个界面，将用户需求标引出的概念图 CG_Q，由此导引下将搜索引擎所返回的摘录结果标引成概念图 CG_S，然后概念图 CG_Q 与概念图 CG_S 两者匹配，按照匹配的准确度，将搜索引擎的检索结果再分类：准确解、近似解、噪声三类。概念图是树形结构，摊平后呈序偶组，于是匹配就归结到序偶组匹配，可仿照生物计算方法。因实时计算需要，概念图标引、匹配还可按实用要求进行简化、规约，有效地降低计算复杂度。因此，作为应用研究探索，后续工作是设想开发一个基于概念图的信息二次检索系统原型。

参 考 文 献

[1] John. R. Taylor. 《语言的范畴化：语言学理论中的类典型》.外语教学与研究出版社. 2001.

[2] 郭绍虞. 《同义词词林》序. 上海辞书出版社. 1982.

[3] 王力. 《使用解字组词词典》序. 上海辞书出版社. 1986.

[4] Hui Liu, Jinglei Zhao, and Ruzhan Lu. An Example-based Approach to the Semantic Analysis of Questions, Journal of Computational Information Systems, 2008, 4(4).

[5] Hui Liu, Jinglei Zhao, Maosheng Zhong, and Ruzhan Lu. Two Phase Semantic Analysis of Real World User Queries, Journal of Computational Information Systems, Accepted for publication in 2009.

[6] Yi Hu, Ruzhan Lu, Yuquan Chen and Hui Liu. A New Hierarchical Conceptual Graph Formalism Adapted for Chinese Text Retrieval. In: Proceedings of IEEE FSKD 2007, Vol.2.

[7] 胡熠，陆汝占，刘慧. 面向信息检索的概念关系自动构建. 中文信息学报, 2007, 21(5).

[8] DUAN Jianyong, LU Ruzhan. A Bio-inspired Approach for Multi-Word expression Extraction. Proceedings of the 21st International Conference on Computational Linguistics AND 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006). Sydney.

[9] Duan,Jianyong Li Ru, Hu Yi. A bio-inspired application of natural language processing: A case study in extracting multiword expression. Expert Systems with Applications, 36(2009):4876-4883

[10] Hui Liu, Jinglei Zhao, Ruzhan Lu. Model Checking a Rule-based Parser, in Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, 2007(NLP-KE 2007), Beijing, China

[11] Hui Liu, Jinglei Zhao, and Ruzhan Lu. Towards the Formal Verification of a Unification System, IEEE Transactions on System, Man and Cybernetics: Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 2009, 39(2):399-408

[12] 方华, 陆汝占,刘绍明. 一个实现多种切分标注算法的系统. 计算机工程, 2004, 30(024):122 - 124.

[13] 樊玉俊 胡熠 陆汝占.基于机器可读词典的词汇知识抽取.计算机应用与软件.2008,25(6)

[14] 宋孜攀,陆汝占.机器可读词典中词汇属性信息的获取.计算机工程与应用.2009,45(5): 138-140

[15] Yi Hu, Ruzhan Lu, Yuquan Chen, Jinglei Zhao. Automated Extraction of Conceptual Knowledge from a Chinese Machine-Readable Dictionary. In: Proceedings of IEEE FSKD 2007, Vol. 4: pp. 578-582.

[16] T. Berners-Lee. What the semantic web isn't but can represent, 1998.