



面向智能搜索的动态知识网络建模^{*}

刘 剑^{1,2,3}, 许洪波¹, 贾岩涛¹, 程学旗¹

(1. 中国科学院计算技术研究所网络数据科学与技术重点实验室 北京 100190;

2. 中国科学院大学 北京 100190; 3. 解放军外国语学院语言工程系 洛阳 471003)

摘 要:随着互联网数据的爆炸式增长和网民获取信息需求的不断增强,传统的搜索方式在移动搜索领域已经难以满足用户的需求,迫切需要将搜索方式从基于词层面提高到基于语义层面,实现基于语义理解的智能搜索。面向开放的互联网数据资源,提出了“动态知识网络+计算算子”的智能搜索模式。在此基础上,详细阐述了动态知识网络的理论基础、结构模式、系统模型及其特点,并且给出了基于动态知识网络支撑智能搜索的基本结构框架,从而对面向语义理解的智能搜索提供理论和模型支撑。最后,对未来研究过程中面临的主要问题和挑战进行了展望。

关键词:智能搜索;知识网络;超图;语义理解

doi: 10.3969/j.issn.1000-0801.2014.10.011

Dynamic Knowledge Network Modeling Orient to Intelligent Search

Liu Jian^{1,2,3}, Xu Hongbo¹, Jia Yantao¹, Cheng Xueqi¹

(1. Key Laboratory of Web Data Science & Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100190, China;

3. Department of Language Engineering, PLA University of Foreign Languages, Luoyang 471003, China)

Abstract: With the explosive growth of internet data and the increasing user demand, the traditional search methods has been difficult to meet user's needs in the mobile search field. In order to achieve the intelligent search based on semantic understanding, it is urgent to improve the level of search from word-based to semantic-based. Facing to the open internet data resources, an intelligent search pattern on “dynamic knowledge network + calculation operator” was proposed. Then, dynamic knowledge network was described in detail, which included theoretical basis, structure model, system model and features. Furthermore, it was presented about the basic structure of intelligent search based on dynamic knowledge network, which provided the support about theory and model for intelligent search. Finally, the main problems and challenges in the future were explained.

Key words: intelligent search, knowledge network, hyper-graph, semantic understanding

1 引言

近年来,随着移动互联网的高速发展,传统基于个人电脑的上网方式正在加速向基于智能移动终端的方式转

变,移动互联网已经成为人们获取信息的主要途径。来自中国互联网络信息中心(CNNIC)的报告显示^[1],截至2013年12月,中国搜索引擎用户规模达到4.9亿户,手机搜索网民数达到3.65亿人,移动式搜索成为不断兴起的

^{*} 国家重点基础研究发展计划(“973”计划)基金资助项目(No.2012CB316303, No.2013CB329602),国家自然科学基金资助项目(No.60933005, No.61232010, No.61173064),国家科技支撑计划基金资助项目(No.2012BAH39B04)

新型应用之一。与此同时,网络空间(cyberspace)中各类应用的层出不穷引发了数据规模的爆炸式增长,形成了网络空间的大数据^[2]。随着互联网数据的爆炸式增长和网民获取信息需求的不断增强,传统的“关键词”搜索局限性逐渐暴露,基于互联网的海量搜索在移动搜索领域已经难以满足用户需求,用户对于搜索结果的直接性要求体现得很明显,精准信息才是移动式搜索用户最想要的。在整合海量互联网碎片化信息的基础上,如何基于用户的片段输入准确理解用户搜索意图,然后从海量显性和隐性知识资源中按照人们需求,有针对性地提炼知识内容或问题解决方案,从而以直接给出可能答案或者更为丰富语义关联信息的形式返回给用户,这是移动搜索所面临的巨大挑战。

传统 Web 资源中的语义信息以自由文本的方式存在,缺乏机器可理解的语义,搜索引擎难以自动有效地整合这些数据,同时资源间的语义关系以一种隐含的方式存在,这些语义信息由于缺乏明确的描述而丢失。因此,对于搜索引擎而言,准确理解数据符号背后所包含的语义信息变得至关重要。为了能够将搜索结果准确地传递给用户,需要引入语义技术,对搜索结果进行优化计算,从而理解用户的搜索意图。为了解决语义缺失问题,互联网创始人 Lee T B 在 XML2000 国际会议上正式提出语义 Web 的体系框架^[3],希望使网络中的信息具有语义,以便计算机能够自动地处理和理解数据。语义 Web 中“语义”的核心就是知识共享,知识共享实质上是基于语义技术的共享,而基于语义技术的智能搜索使得搜索引擎不再拘泥于用户所输入请求语句的字面本身,而是透过现象看本质,准确地捕捉到用户所输入语句后面的真正意图,并以此进行搜索,从而更准确地向用户返回最符合其需求的搜索结果。

在 2013 年 5 月的 Google I/O 大会上,Google 的 Amit Singhal 提出了未来搜索引擎的设想:搜索引擎的 3 个主要功能需要改进,搜索将需要答案、对话、预测。未来的搜索引擎需要更智能地为用户服务,这一切离不开富含语义信息的知识库作为基础支撑。语义 Web 希望赋予互联网上所有资源唯一的标识,以一种明确、形式化的方式描述信息资源,从而在资源之间建立起机器可以处理的各类语义关联,最终将万维网中现存的信息发展成一个巨大的全球语义知识库^[4]。但是,面对海量的数据资源、丰富的文档类型、形态各异的数据格式,数据资源的耦合度较低,也缺乏统一的管理,难以形成统一的语义知识库。因此,基于现有的互联网数据资源,构建大型的语义知识库,为智能搜索

提供语义知识支持成为切实可行的方案。本文面向开放的互联网数据资源,结合现有技术应用,提出以超图(hyper-graph)理论为基础的动态知识网络(dynamic knowledge network, DKN)建模方式,从模型层面阐述了“知识+计算→智能”的智能搜索模式,通过计算算子实现基于知识的计算,从而对面向语义的智能搜索提供理论支持和模型支撑。

本文首先结合信息技术的发展介绍了互联网搜索技术的现状,分析了基于语义技术进行智能搜索的发展前景。在此基础上,提出以超图理论为基础、以计算算子为技术支撑,进行世界知识建模的动态知识网络建模方案,阐述了该模型的理论基础、结构模式、系统模型及其特点以及基于动态知识网络支撑智能搜索的基本结构框架,最后,对今后研究工作中所面临的主要问题和挑战进行了展望。

2 研究现状

随着计算机与通信技术的迅速发展,互联网上的信息呈现指数型增长,在互联网信息越来越丰富、用户使用方式也越来越多变的同时,庞大并且关联的信息让大部分用户感到无所适从,搜索的价值也就越来越明显。面对海量信息,基于分类目录和关键词的搜索方式越来越难以适应用户的搜索需求,迫切需要将检索方式从基于词层面提高到基于语义层面,实现基于语义理解的智能搜索。智能搜索不仅要求提升检索技术,还向着信息服务的智能化、个性化、可互动的方向发展,因此,需要实现查询请求和目标资源的语义理解。本体作为知识的承载者被信息科学领域引入,并作为语义 Web 的核心技术,对网络信息资源进行语义表达和标注。根据本体技术在搜索引擎中的作用,将目前的智能搜索划分为 3 类^[5],具体介绍如下。

- 基于传统搜索的增强型搜索:这一类搜索的核心还是传统的搜索引擎,通过本体技术对用户查询词的处理来提高搜索效果,如 IBM 与苹果公司等合作开发的 OntoSeek 系统^[6]、美国斯坦福大学与 IBM 等研究机构联合开发的 Tap 系统^[7]等,还有研究将 wordnet 作为查询扩展和约束,以改善搜索的效果^[8]。
- 基于本体推理的知识型搜索:这一类搜索是基于构建的本体知识库,通过本体推理技术实现知识的自动发现和关联,如美国马里兰大学开发的 SHOE 系统^[9]、上海交通大学提出的 SPARK^[10]、清华大学提出的细粒度语义网络检索模型^[11]等。



· 其他类型的搜索: 还有一些其他类型的搜索模型, 如美国华盛顿大学开发的 KnowItAll 系统^[12]、华中科技大学提出的应用在安全访问控制领域的搜索模型^[13]、上海交通大学与香港科技大学共同提出的一种增强的语义搜索模型^[14]等。特别值得一提的是, Wolfram 在 2009 年发布了 WolframAlpha 系统, 该系统一经发布就引起很大的反响, 甚至有人认为它会取代 Google 的搜索霸主地位。

近年来, 随着 Linking Open Data 等项目的全面展开, 语义数据源的数量激增。互联网正从仅包含网页与网页间超链接的文档万维网(document Web)转变为包含描述各种实体与实体之间丰富关联的数据万维网(data Web)。在此背景下, 谷歌、微软、百度和搜狗等搜索引擎公司纷纷以此为基础构建知识图谱, 分别为 Knowledge Graph、Probase、知心和知立方, 以此来改进搜索质量, 从而拉开了智能搜索的序幕。

3 动态知识网络建模

网络时代人们在探讨数据、信息、知识之间的相互关系时, 认识到数据是事物属性及其相互关系等的抽象表示, 信息则是有目的、有意义、有用途的数据, 而知识是通过信息深度加工, 经过逻辑或非逻辑思维, 认识事物本质而形成的经验与结论^[15]。互联网蕴含着丰富的知识资源, 不论是信息直接所包含的知识还是信息背后所隐含的知识, 都反映在网络中。维娜·艾莉^[16]曾经指出, “我们可以把自己的个人知识看成一张认识的‘网’, 很多想法、感觉、思想、概念和信仰都在这里交织在一起”。因此, 可以从知识的这种网状结构特征得到启发, 互联网中也包含着一张巨大的知识网络, 经过数据的采集和清洗、信息的提炼和抽取、知识的描述和集成, 最终可以利用一个开放的知识网络将其呈现出来。基于这一思想, 本文提出“知识+计算→智能”的智能搜索模式, 如图1所示。

其基本思想是: 基于互联网的各类数据资源, 构建动

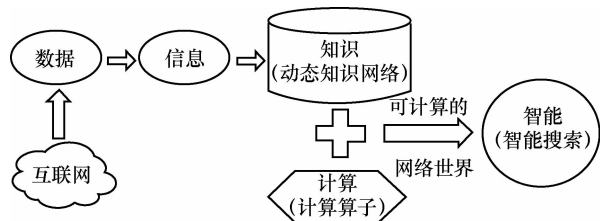


图1 智能搜索模式的设想

态知识网络, 以此为基础, 结合定义好的各类计算算子的灵活组合, 响应用户需求, 实现对用户真实查询意图的语义理解和问题解答, 从而实现通过一个事实知识库和一系列计算算子得到一个可计算的网络世界的构想。

3.1 动态知识网络的理论基础

关于“知识网络”这个概念, 最早是由瑞典工业界在 20 世纪 90 年代提出来的^[17], 不同时代、不同学科和不同领域的学者, 对“知识网络”概念的内涵和外延有着不同的认识^[18]。王众托院士通过对无处不在的网络社会的分析^[15], 认为一个知识网络应该有 3 个层次: 技术层面的技术网络、知识资源内部联系的知识网络、知识在人际间传播的人际关系网络。依据此理论, 本文主要研究知识资源自身存在内在关联的知识网络。知识网络目前还没有明确的定义, 它是一个集合概念, 指的是知识的空间结构集合, 即它是由知识节点和知识关联所构成的集合。其中, 知识节点(以下简称节点)一般代表知识单元的存储单位, 由概念或者事物组成; 知识关联(以下简称边)可分为内部关联和外部关联。内部关联构成知识个体, 表达知识的内涵联系, 外部关联是知识个体之间的外延联系, 构成知识网络的各种链接关系。因此, 知识网络是由节点和边构成的网状结构, 这种网状结构的表现形式多种多样, 有树型结构、星状结构、环型结构、单向关系网络、多向交叉复合关系网络等。

典型的知识网络主要考虑知识节点之间的二元关联, 即两个知识节点之间的关联。通常, 事物之间的关联不仅仅是单一和单向的简单关系, 而是一个复杂和多向的网络。鉴于网络世界中知识节点之间关联的复杂性, 一般的二元网络图难以完全刻画网络世界中知识的特征, 因此, 出现了超越一般网络的网络系统问题。本文研究的知识网络规模巨大、连接复杂, 知识节点具有异质性, 可以称为超网络(hyper-network), 本文用超图来定义该类超网络^[19]。超图这一概念是 Berge 在 1970 年提出的^[20], 超图不同于一般图论中的无向或有向图, 后者的每一个边只连接两个节点, 而超图中的边可以连接两个以上的节点, 称为超边。因此, 本文提出的动态知识网络的模型是用超图表示的超网络。下面给出超图在数学上的严格定义, 见定义 1。

定义 1 设 $V=\{v_1, v_2, \dots, v_n\}$ 是一个有限集, 若满足以下条件: $e_i \neq \emptyset (i=1, 2, \dots, m)$; $\bigcup_{i=1}^m e_i = V$, 则称二元关系 $H=(E, V)$ 为超图。 $V=\{v_1, v_2, \dots, v_n\}$ 是超图的顶点集, $E=\{e_1, e_2, \dots, e_m\}$ 是超图的边集, 集合 $e_i=\{v_{i1}, v_{i2}, \dots, v_{ij}\} (i=1, 2, \dots, m)$ 为超图的边。如

果在超图的边集中定义了方向,那么超图就是有向超图;反之,则是无向超图。

定义 2 (超路径)超图 H 中的顶点和超边交错序列 $\{v_1, E_1, v_2, E_2, \dots, E_q, v_{q+1}\}$ 称为具有长度为 q 的超路径,若满足以下条件:

- $\{v_1, v_2, \dots, v_{q+1}\}$ 在超图 H 中彼此不同;
- $\{E_1, E_2, \dots, E_q\}$ 在超图 H 中彼此不同;
- $v_k, v_{k+1} \in E_k, k=1, 2, \dots, q$ 。

同时 $q \neq 1$ 且 $v_{q+1}=v_1$,则这一条超路径称为长度为 q 的超回路。

超图是对图的一种扩展,其在描述多个节点之间拥有复杂多元关系的动态知识网络时,具有极大的优势。比如有 8 个知识节点 $V=\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$, 构成 4 个多元关系 $E=\{e_1, e_2, e_3, e_4\}$, 其中, $e_1=\{v_1, v_2, v_4\}$, $e_2=\{v_2, v_3, v_4\}$, $e_3=\{v_4, v_5, v_8\}$, $e_4=\{v_5, v_6, v_7, v_8\}$, 用超边表示多元关系,可以得到如图 2 所示的超图。

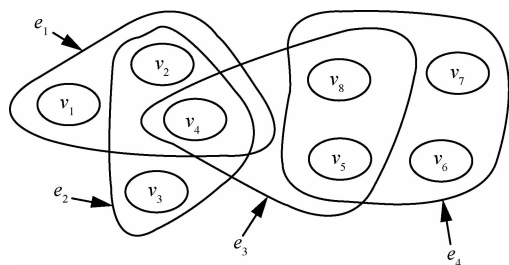


图 2 超图表示的多元关系

3.2 动态知识网络的系统建模方案

3.2.1 动态知识网络的结构模式

动态知识网络是对互联网域空间知识的描述,是表示知识节点及节点间相互关联的复杂网络系统。本文基于超图理论对动态知识网络进行建模,拓展了普通图中的节点和关系的类型,能够更加灵活地实现知识的概念化描述。为了实现知识的语义表达,需要从结构上对其模式进行分析,解释知识网络的模式是如何由一些简单的子模式(模式基元)

组合而成的。表 1 为动态知识网络的结构子模式示例。

通过对节点和关系描述的拓展,动态知识网络能够描述更加复杂的结构,也使得对子模式的提取具有更加丰富的语义信息。结合超点和超边子模式,还可以衍生更加抽象和复杂的子模式。在一些应用中,通过这些子模式进行动态知识网络的分解和缩减可以简化结构的复杂性,从而在更高层次上分析网络结构的特性。

3.2.2 动态知识网络的系统模型

面向开放网络数据环境,本文结合相关研究^[20],提出动态知识网络的系统模型,针对海量数据中知识的各种特征表现和复杂关联进行语义知识表达和操作。动态知识网络的系统模型使用一个七元组表示,即 $DKN=(V, E, A, Val, F, G, O)$,其中, V 是知识节点的非空有限集合, E 是知识关联的非空有限集合, A 是知识节点和知识关联属性的非空有限集合, Val 是属性的值域集, F 是知识节点和知识关联上的属性值映射函数集, G 是知识节点上的关联映射函数集, O 是针对知识网络的各类操作,即计算算子的集合。下面分别对该模型的组成元素进行介绍。

(1) 知识节点 V

知识节点由在认识上具有独立性的知识元素构成,具有层次性,其最小粒度可以称为知识元,是独立不可再分的知识元素,如人名、城市名等。知识元是最小的知识节点,多个知识元通过知识关联可以构成更大的知识节点,知识节点的集合可以构成知识体系。对于 $V=\{v_1, v_2, \dots, v_n\}$, 知识节点 v_i 代表一个简单或者复杂的事物或概念。

(2) 知识关联 E

知识关联是构成动态知识网络的知识节点之间的关联关系,这种关联表现为以一种拓扑形式存在的网络结构,其网络性体现在知识因为本身的某种联系而相互聚集形成网络。 $E=\{e_1, e_2, \dots, e_n\}$ 是带有标签的有向超边和无向超边的集合,超边 e_i 代表一个简单或者复杂的知识关联,其最小粒度是独立不可再分的关联关系。有向超边 $e_i=\langle r_i \rangle$,

表 1 动态知识网络的结构子模式示例

图形示例	基元类型	语义描述	形式化描述
	点基元	实体特征	v_i , 无嵌套、无关系、独立实体
	超点基元	嵌套特征	v_i^s , S 为嵌套层次, $S=0$ 时为点基元
	超边基元	关系特征	e_j , 带标签的超边, $ e_j \geq 2$
	路径基元	解特征	$path()=\{v_1, e_1, v_2, e_2, \dots, e_q, v_{q+1}\}$
	星状基元	度特征	$DKN'(x)$ 为 DKN 中所有包含点 x 的超边所导出的部分超图
	生成超树基元	支撑特征	生成超树 OKN_i 为 OKN 的一个生成子超图



$(\lambda_i)>$ 是一个序偶, r_i 是 e_i 中输入变量的集合, λ_i 是 e_i 中输出变量的集合;无向超边 $e_i=\{v_1, v_2, \dots, v_m\}$ 是一个多元无向边集合。通常有 3 种基本的知识关联类型:同一性关联,知识节点间具有某种共同性质形成的关联,主要表现为知识节点间的继承性,知识节点的等同性是同一关联的特殊表现;隶属性关联,构成知识节点的单个知识元或者知识元集合隶属某个概念、类别和范畴的逻辑关系,主要表现为知识节点间的属性关系、分类关系、包含关系等;相关性关联,是在同一性关联、隶属性关联之外的,知识节点间大多具有的相互依存、相互作用的关联,主要表现为工作、家庭、应用、影响等各种关系,这种关系不是严格固定的,其数量关系也是不完全确定的,它使得知识节点间在横向上形成关联网络。

(3)属性 A 、属性的值域集 Val 、属性值映射函数 F

一个具体的事物或者概念总是通过一些性质加以描述和区分,属性用来描述知识节点和知识关联自身的性质和特征。具体又可以将属性分为数值型属性和对象型属性两大类。 $A=A_V \cup A_E$,其中, A_V 是知识节点属性集, A_E 是知识关联属性集。 $Val=Val_V \cup Val_E$,是知识节点属性和知识关联属性的值域集合。 $F=F_V \cup F_E$,是知识节点、知识关联与各自属性值的映射函数,其中, $F_V: V \times A_V \rightarrow Val_V$, $F_E: E \times A_E \rightarrow Val_E$ 。比如 $V=\{v_1, v_2, v_3\}$ 代表 3 个企业,组成的商业合作可以用一条超边 $e_1=\{v_1, v_2, v_3\}$ 来描述,其中 A_V 包含“公司名称”、“成立时间”、“年营业额”等企业的属性, A_E 包含该商业合作的属性,如“组建时间”、“合作模式”等,属性取值 Val_V 包括“中远集运”、“中海集运”、“中国外运”等, Val_E 包括“2014 年 5 月”和“合作经营”等。在动态知识网络系统中,不同属性的对应取值及取值约束通过 F 来确定,如 $F(\text{组建时间}) \rightarrow \text{Date}(\text{2014 年 5 月})$ 。另外,利用属性值映射函数还可以实现节点间基于属性的聚类。

(4)知识关联映射函数 G

动态知识网络中各个知识节点按照需要的因素、层次、结构和功能等构成结构化的知识网络系统,这种结构化的过程需要将知识节点通过某些方式关联起来,即知识关联映射。 G 是 V 上的关联函数集合, $G=\{g|g(v)=e\}$,表示事物之间的不同关联类型。当然,也可以理解为 G 是超边构造函数,反映了知识节点间超边的构造关系。构建知识节点间的知识关联是建立动态知识网络的关键环节, G 决定了哪些知识节点在同一个知识关联中以及知识关联是如何划分的。根据知识节点间关联类型是否明确,可以有两种构建方式:对于可以预定义的关联模式,采用先知识

关联后知识节点的构建方式,即先确定动态知识网络的关联模式,描述为超边,然后在知识节点集中搜索符合各关联模式的知识节点对;对于关联模式比较模糊的情况,可以通过对知识节点进行信息抽取,寻找它们之间的关联模式,常用聚类、频繁项集挖掘等方法。

(5)计算算子 O

计算算子主要针对知识网络完成各类运算操作,如同普通运算符号作用于数后,可以得到新的数,一个算子作用于一个输入后,可以实现从一个知识网络空间到另一个知识网络空间(或它自身)的映射。根据实际需求,将算子分为两大类:一是实现动态知识网络内部元素动态构造的构建类算子,二是提供外部服务的应用类算子。基于外部信息输入的计算算子模型如图 3 所示。

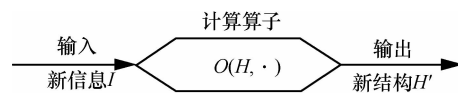


图 3 计算算子模型

基于外部信息输入的计算算子模型可以表示为: $H' \cong O(H, \cdot)I$,其中, $O(H, \cdot)$ 为计算算子, I 为新输入的信息, H 为已有知识网络图或子图结构, H' 为经过算子运算后输出的新知识网络图或子图结构, $\cong$ 表示输出的新结构 H' 是外部信息 I 在已有结构 H 上经过计算算子 $O(H, \cdot)$ 运算后产生的结果。

结合新信息的输入,计算算子封装一些针对动态知识网络常用操作的灵活组合,从而实现基于计算算子的运算。表 2 给出了一些针对动态知识网络的常用算子示例。

3.2.3 动态知识网络模型的特点

本文通过对网络世界知识进行建模,提出基于超图理论的动态知识网络系统模型,该模型具有以下几个方面的特点。

(1)可以描述复杂知识节点和知识关联

动态知识网络模型中允许定义复杂知识节点,知识节点和知识关联在一定程度上可以相互转化,因此,复杂知识节点可以是多个知识节点、知识关联的集合。知识关联复杂多样,既有二元关联,又有多元关联。既有明确定义的关联类型,又有难以明确描述的关联类型,因此具有很强的知识描述能力,其完整形态是一个多元、异构、立体的超网络。

(2)结构开放、灵活,可扩展性强

动态知识网络模型描述的知识是可扩展的,可动态感知数据的变化,同时具有时效性,随着新信息的加入而动态更新。另外,模型中允许定义新的知识关联,通过关联结构的可变实现网络结构的灵活性,同时对未知关联类型的

表 2 动态知识网络的常用算子示例

算子名称	形式化描述	含义解释
节点添加	$add_vertex(v_i, E_j)$	点基元添加需指定包含该节点的关系集合
属性聚类	$clustering(V, A)$	对点基元集合基于属性的聚类
求两点间超路	$path(v_i, v_j, h)$	提供两个点基元间的可能超树基元
求两点间最短超路	$short_path(v_i, v_j, h)$	提供两个点基元间的最短路径基元
求路径的数量	$calculate_path(v_i, v_j)$	提供两个点基元间的可能路径基元数

包容性使得面对不确定的环境时,网络结构也可以随着信息的交互而发生演化。

(3)具有处理不确定、不精确信息的能力

动态知识网络模型是基于超图理论的,而超图中的集合理论是其核心,因此,基于集合的表达方式适合描述非明确定义的关系和规则。对于难以被明确定义和精确描述的知识关联,该模型采用无向超边进行描述,同时,模型中定义的计算算子可以实现对知识网络的各类操作,使得能够利用图理论来处理网络环境下不确定、不精确的信息。

(4)具有较强的可计算性

基于图理论,动态知识网络模型中定义了多种类型的图操作,通过这些图操作的灵活组合,计算算子可以实现动态知识网络构建和应用过程中的模式识别、路径分析、子模式构建等各种计算功能。另外,在特定的应用需求驱动下,基于一定的规则和约束条件,还可以进行知识的推理计算。

4 基于动态知识网络的智能搜索

新一代的智能搜索致力于建立一个智能化、个性化和

互动的搜索模式,这就需要对查询请求的语义理解和对目标文档的语义理解,而语义理解是建立在语义知识库基础之上的。本文提出动态知识网络的系统建模方法,通过构建动态知识网络,提高搜索引擎的语义理解能力,从而达到智能搜索的目标。基于动态知识网络的智能搜索框架如图 4 所示。

其基本思想是:充分利用现有各类数据资源,基于动态知识网络系统模型,结合机器学习和数据挖掘等技术,构建动态知识网络,并以此为语义基础,支持基于语义理解的智能搜索。动态知识网络对智能检索的语义支持通常包含以下两个方面。

(1)语义的扩展与优化

用户以自然语言输入查询,系统基于动态知识网络进行语义分析,需要理解用户提交关键词搜索背后的真正意图,主要包括分类、属性、同义等语义关系的提取、歧义消解等,从而丰富查询的语义信息。同时,在语义理解的基础上,对数据资源进行整合处理,获取真正符合语义的信息资源。

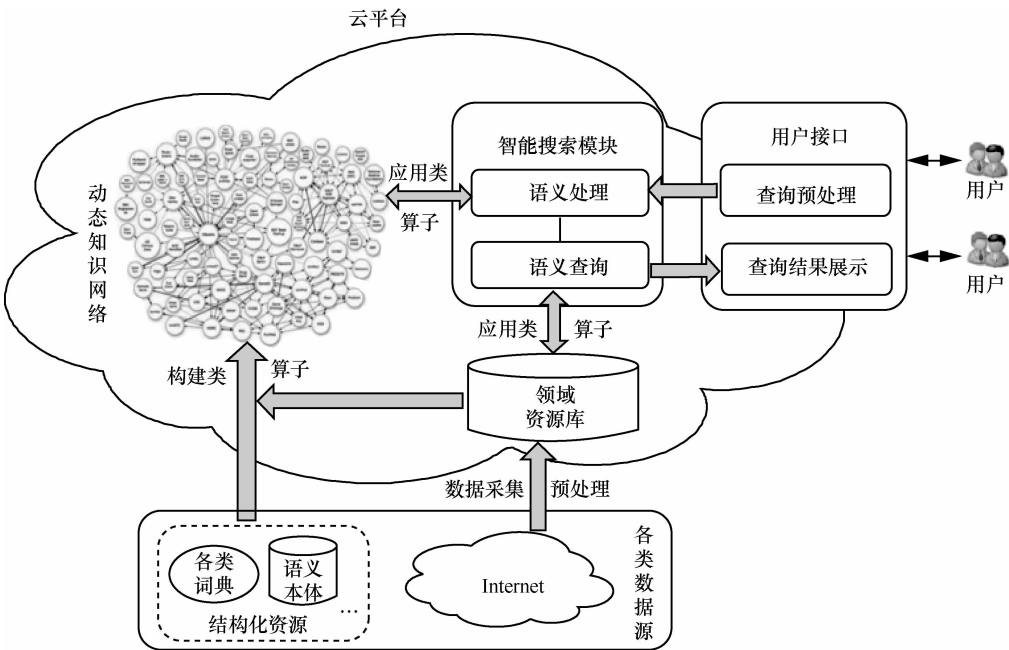


图 4 基于动态知识网络的智能搜索框架



(2)语义的推理与计算

基于动态知识网络,对用户查询的关键词进行概念化或者实例化处理、相似性计算等,从而在更高层次或者更细粒度上建立语义关联。同时,根据动态知识网络的推理规则,进一步拓展语义的关联和约束。

5 结束语

本文通过对现有信息检索和语义处理技术的介绍,分析了基于语义技术进行智能搜索的发展前景。基于此,提出以超图理论为基础、以计算算子为技术支撑进行网络世界知识建模的动态知识网络建模方法,并详细阐述了该模型的理论基础、结构模式、系统模型及其特点,最后给出了基于动态知识网络支撑智能搜索的基本结构框架,为基于语义的智能搜索应用提供了有效的模型和方法支持。尽管目前已经进行了一些探索性的研究工作,但未来的工作仍然面临两个重要挑战:动态知识网络系统模型的完善;动态知识网络支撑的智能检索应用。

参考文献

- 1 中国互联网络信息中心. 2013 年中国搜索引擎市场研究报告, 2013
- 2 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 2012, 27(6): 647~657
- 3 Lee T B. Semantic web on XML. <http://www.w3.org/2000/talks/1206-xml2k-tbl>, 2014
- 4 王本年, 高阳, 陈世福等. Web 智能研究现状与发展趋势. 计算机研究与发展, 2005, 42(5): 721~727
- 5 文坤梅, 卢正鼎, 孙小林等. 语义搜索研究综述. 计算机科学, 2008, 35(5): 1~4
- 6 Guarino N. Ontoseek: content-based access to the web. IEEE Intelligent Systems, 1999, 5(6):70~80
- 7 Guha R, McCool R. TAP: a semantic web test-bed. Journal of Web Semantics, 2003,1(1): 81~87
- 8 Kruse P M, Naujoks A, Roesner D, *et al.* Clever search: a wordnet based wrapper for internet search engines. Proceedings of the 2nd GermaNet Workshop, Bonn, Germany, 2005:367~380
- 9 Heflin J, Hendler J. Searching the web with shoe. Proceedings of AAAI-2000 Workshop on AI for Web Search, Austin, Texas, 2000:450~455
- 10 周琦. 基于关键词的语义搜索. 上海交通大学硕士学位论文, 2009
- 11 吴刚, 唐杰, 李涓子等. 细粒度语义网检索. 清华大学学报(自然科学版), 2005, 45(1): 1865~1872
- 12 Cafarella M J, Downey D, Soderland S, *et al.* KnowItAll: fast, scalable information extraction from the web. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vancouver B C, Canada, 2005: 563~570
- 13 文坤梅. 基于本体知识库推理的语义搜索研究. 华中科技大学博士学位论文, 2007
- 14 Zhang L, Yu Y, Zhou J, *et al.* An enhanced model for searching in semantic portals. Proceedings of the International Conference on World Wide Web, Chiba, Japan, 2005:453~462
- 15 王众托. 无处不在的网络社会中的知识网络. 信息系统学报, 2007, 1(1): 1~7
- 16 赵蓉英. 论知识网络的结构. 图书情报工作, 2007, 51(9): 6~10
- 17 田占伟, 张庆普, 刘臣. 语义知识网络的结构分析与构建. 情报理论与实践, 2011, 34(10): 113~118
- 18 刘向, 马费成, 王晓光. 知识网络的结构及过程模型. 系统工程理论与实践, 2013, 33(7): 1836~1844
- 19 王志平, 王众托. 超网络理论及其应用. 北京: 科学出版社, 2008
- 20 王众托. 关于超网络的一点思考. 上海理工大学学报, 2011, 33(3): 229~237
- 21 吴颖敏. 市场机遇发现的超图支持方法研究. 华中科技大学博士学位论文, 2009

[作者简介]



刘剑,男,中国科学院大学博士生,解放军外国语学院讲师,主要研究方向为知识工程、数据挖掘。



许洪波,男,博士,中国科学院计算技术研究所副研究员,主要研究方向为数据挖掘、信息安全。

贾岩涛,男,博士,中国科学院计算技术研究所助理研究员,主要研究方向为社会计算、信息安全。

程学旗,男,博士,中国科学院计算技术研究所研究员,主要研究方向为互联网搜索与挖掘、网络科学与社会计算、网络与信息安全。

(收稿日期:2014-08-11)