基于对象网格的网络科技信息重要对象识别方法研究 1)

邹益民 [1] 张智雄 [2]

[1](浙江师范大学经济与管理学院，浙江金华 321004)

[2](中国科学院国家科学图书馆，北京 100190)

摘要 利用网络科技信息支持科技动态监测和战略决策分析日益成为情报机构的一项重要工作，但面对良莠不齐的海量网络科技信息，如何快速而准确的对语篇论述的主题进行揭示，则是一个亟待解决的重要问题。本文通过把语篇映射成一个蕴含对象语法信息、语义信息、位置信息、共现信息、分布信息以及语篇结构信息的对象网格，将非结构化的语篇转变为可计算的知识单元。根据对象在网格中的分布规律，以及由这些规律凝练的对象凝聚度、活跃度和生命跨度三个指标维度，对语篇中重要的对象进行识别。利用具有明确概念的知识对象对语篇中蕴含的重要情报线索进行揭示。

关键词 知识对象；对象网格；对象计算；语篇表示

分类号 G250

Identify the Important Objects of the Web Resource Based on Knowledge Object Grid

ZOU Yimin[1] ZHANG Zhixiong[2]

[1] College of Economics and Management, Zhejiang Normal University, Jinhua 321004)

[2](National Science Library, Chinese Academy of Sciences, Beijing 100190)

Abstract Nowadays, using web scientific information to support the analysis of dynamic monitoring and strategic decision-making is becoming an important task for intelligence analysis teams. How to correctly and quickly identify the core topic of a discourse from a large number of gathered web resources has become an urgent issue. In this paper, A new discourse representation model – object grid is used to map unstructured discourse into computable knowledge unit, which contains the information about object's grammatical role, semantic information, position information, co-occurrence information, distribution information, and discourse structure information. According to the distribution laws of objects in the grid, three dimensions of indicators about object's cohesion, activity and life span is used to identify important objects. Using explicit concepts reveal important intelligence clues inherent in the discourse.

Keywords Knowledge Object ; Object Grid; Object-based Computing; Discourse Representation

1 引言

在网络日益成为最重要的科学交流和传播渠道的今天，很多重要的科技战略、科研活动、科研成果都是通过网络实现信息的对外发布。例如：美国奥巴马政府 2011 年国家创新战略[1]和 2012 年的大数据计划[2]、欧盟的创新积分榜[3]、世界经合组织的主要科学和技术指标[4]等与科学研究和科技战略决策相关的信息都可以直接通过网络获取。战略情报研究团队可以通过这些资源对科技发展重大趋势及战略倾向进行把握，及时发现领域内的重大科技问题和重大研究进展。但面对良莠不齐的海量网络信息，如何准确而快速的对网络科技信息中蕴含

作者简介：**邹益民**，1983 年生，情报学博士，主要研究方向：网络数据挖掘、智能信息处理与信息系统等，E-mail：yimzou@gmail.com。**张智雄**，1971 年生，研究馆员，博士生导师，博士，主要研究方向：知识抽取、智能信息处理、知识组织等, E-mail：zhangzhx@mail.las.ac.cn。

1) 本文系国家自然科学基金“基于语言网络的文本主题中心度计算方法研究”（项目编号：61075047）的研究成果之一。

的情报线索进行捕获，则是网络科技信息自动监测研究面临的一个重要问题。在对网络科技信息的研究中，笔者发现很多的重要术语和命名实体，例如：战略计划、科研人员、研究报告、战略声明、R&D 投入、重大项目和法案、科研资助组织等，一起内嵌在科学创新机构站点发布的网页、RSS 和富文档当中。笔者将这些命名实体和术语称为知识对象或对象[5]。这些对象具有很强的概括能力，能够表达和传递明确的概念，并且携带了网络资源的关键信息。通过对这些网络科技信息中蕴含的重要对象进行识别，能够很好的对网络科技信息论述的中心主题进行揭示。基于此，本文提出了基于对象网格的网络科技信息重要对象识别算法，以单篇网络科技信息为研究对象，利用对象网格对网络科技信息进行表示，通过对象在网格中的分布规律完成网络科技信息中重要对象的识别，并对该方法的实验效果进行了分析。

2 相关研究

网络科技信息资源通常有两部分组成，一部分是网络信息的主体内容，例如：新闻网页中新闻内容部分，它是网络信息的核心；而另外一部分则是与网络信息论述“主题”无关的导航条、广告信息、版权信息等内容。本文将网络信息资源的主体内容称为语篇，研究的问题也主要围绕语篇中的重要对象识别展开。其与关键词[1]自动抽取的研究类似，相关研究大致可分为有监督和无监督方法。

有监督方法通常是将关键词抽取转化为一个二元分类问题，从训练语料中获取分类模型，来判断一个词是否为关键词，例如：Frank 等[6]和 Witten 等[7]采用朴素贝叶斯、Turney 等[8]采用 C4.5 决策树、Zhang 等[9]采用支持向量机分别进行了关键词的自动抽取。有监督的方法能够达到较高的准确度，其稳定性也较高，但学习模型的构建往往需要大量人工标注的训练集，这不但非常耗费人力，而且在网络环境下常常是不现实的，更重要的是用户关注会随着领域热点变化而发生变化，如果学习模型得不到及时更新，将会影响关键词抽取的效率。

无监督方法又可进一步分为基于统计、特征权重计算和图结构的分析方法。基于统计的方法主要是利用 N-Gram[10]、词频[11]、TFIDF[12]等统计信息获得关键词，这种方法简单易行，通用性强，但基于统计的相关方法孤立的考虑词的频次而忽略了其在文本中的相互影响，并且对低频关键词无法进行揭示，使得抽取关键词准确率并不高；特征权重计算则是通过为词首次或者最后一次出现的位置[13]、词的长度[14]、词性[15]、词是否在标题中出现[16; 17]等特征对关键词进行抽取，这种方法对语篇中词之间的语义关系缺乏考虑，且在权重分配时往往过于主观；图结构的方法通过考虑词之间的关系将语篇映射成为网络图，通过对图的挖掘实现关键词的抽取，这方面的研究包括：Rada Mihalcea 和 Paul Tarau[18]在 PageRank 算法基础上提出的 TextRank，Marina Litvak 等[19]利用 HITS 算法来计算图中词结点的重要度，Huang 等[20]提出通过构建语义结构网络进行关键词抽取，利用小世界作为特征权重，基于同句共现构建无向语言网络等。图结构方法能够实现对低频关键词的识别，但这些图的构建往往是基于词之间的单一关系（语法关系、语义关系或共现关系等）进行的构建，对语篇中词之间存在的多重关系缺乏考虑和融合，另外，图中结点之间的关系往往过于稀疏，不能很好的利

[1]关键词仅包含一个词，而关键词语至少包含两个词，但人们通常把关键词与关键词语统称为关键词，为了对不同的概念进行区分，本文以“关键单词”表示仅包含一个单词的关键词，而以“关键词”表示通常意义上的关键词，既包含关键单词也包括关键短语。

用图的相关理论对其进行挖掘。

3 基于对象网格的语篇表示模型

对网络科技信息进行深层的语义解析，首先需要将网页或者富文档中的主体内容信息（本文称其为语篇）转为一种表示模式，使其能够将半结构化和非结构化的自然语言文本映射为计算机能够识别、处理、分析的结构化特征表示形式。为了尽可能多保留语篇中知识对象的信息，包括：位置信息、语法信息、语义信息、共指信息、分布信息以及语篇的结构信息，笔者在实体网格（Entity Grid）[21-24]的基础上提出了基于对象网格（Object Grid）的语篇表示模型。在对象网格中，语篇被映射成一个由对象及其语法角色组成的网格，对象网格是一个二维的数组，网格的列对应语篇中的句子而行对应语篇中的对象，语篇中的每一个对象在网格中都对应于其在给定句子中的语法角色。在对句子粒度的界定上，由于子句的分割将引入相应的噪声，对象网格将主句作为其分析的基本单元。对象网格用“S”、“O”、“X”和“－”来标识对象在句子中的语法角色，其中“S”对应主语、“O”对应宾语、“X”对应非主语和宾语的其它句法角色、“－”标识相应的对象在给定的句子中不存在，并且对象语法角色具有一定优先级：S>O>X>－，当同一个对象在给定的句子中具有不同的语法角色时，取优先级最高的角色进行标识。图 1 中也语篇内容对应的对象网格如图 2 所示。

1. [Barack Obama]s is [the 44th and current President of the United States]o, in [office]x since 2009.
2. [He] s is the first [African American]o to hold the [office]o .
3. Born in [Honolulu, Hawaii]x, [Obama] s is a [graduate]o of [Columbia University]x and [Harvard Law School]x, where [he]s was [president]o of the [Harvard Law Review] x.
4. [He]s was a [community organizer]o in [Chicago]x before earning his [law degree]o. [He] s worked as a [civil rights attorney]o in [Chicago]x and taught [constitutional law]o at the [University of Chicago Law School]x from 1992 to 2004.

图 1 语篇内容

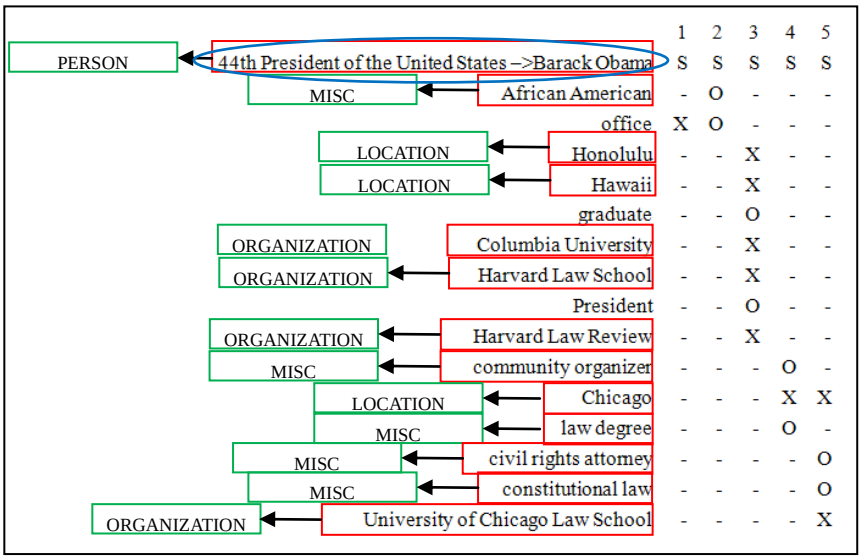

图 2 对象网格

实体网格是麻省理工学院的 Barzilay 和爱丁堡大学的 Lapata 受向心理论（Centering Theory）的启发提出的语篇表示模型[21]，用于对语篇的局部连贯性进行评测。但如果将其应用于语篇重要对象的识别上，还存在很多不足和局限，对象网格对实体网格的扩展主要体现在以下三个方面：

（1）将中心名词扩展成知识对象。在实体网格中用实体（通用名词和命名实体）的中心名词代替实体，但对于那些具有明确含义和类型的命名实体和领域术语来说，用其中心名词来代替，将会使其丧失原有的含义。例如：图 2 中的“Columbia University”在实体网格中将被用“University”来代替。在对象网格中用中心名词被扩展成知识对象，知识对象中包含了命名实体和领域术语等具有明确含义的知识单元，是学科知识的专业用语，与领域知识密切相关，负载着很大的信息量。同时，知识对象也是语篇情报价值的重要线索。因此，将知识对象作为一个整体融入到对象网格中，使得对象网格能够传递更多的语篇信息，也是利用对象网格进行重要对象识别的重要前提。

（2）将中心名词以外的名词融入对象网格。在实体网格中，实体中除中心名词以外的词都将被丢弃，丢失的对象不但使得网格失去了很多语篇信息，也会影响到对实体在网格中实体间的共指信息以及实体分布规律的捕获。本文根据复合名词的特点对其进行拆分，将所有拆分后的对象赋予相同的语法角色，例如：对象“nuclear waste R&D”，被拆分成“nuclear waste”和“R&D”两个对象，并赋予和“nuclear waste R&D”同样的语法角色。拆分后的对象同样具有丰富的语义信息，而且笔者认为“nuclear waste”和“R&D”对对象“nuclear waste R&D”而言，具有同样的语义“贡献度”，对语篇情报价值的判断具有同等重要的作用，所以为拆分后的对象赋予了相同的语法角色。将中心名词以外的名词融入对象网格，不但有利于保留更多的语篇语义信息，也为对象在网格中分布规律的捕获奠定了良好的基础。

（3）对知识对象进行语义合并。在实体网格中一个重要的假设是实体之间是相互独立的，其只对具有指代关系的实体进行合并，而对语义相关的实体则缺乏处理，例如：图 2 中的“Barack Obama”和“44th and current President of the United States”指的是同一个知识对象，如果不对这些语义相同，形式不同的对象进行处理，将会造成很多稀疏的行，影响对象在网格中分布规律的捕获。为了改变网格中实体之间相互独立的假设，有效捕捉到对象之间的词汇凝聚因素，在对象网格中对语义相关的对象进行合并。在本研究中，笔者并不试图去改进对象之间的语义相似度计算算法，来提高语义相关对象的聚合效率，而是只将具有明确相同含义的对象进行合并。另外，针对在同一句子中对象合并的情况，取对象在句子中优先级最高的语法角色用于表示合并后的语法角色，例如：图 2 第一句“Barack Obama”和“44th and current President of the United States”的语法角色分别为“S”和“O”，则取 S 代替。

4 基于网格分布的重要对象识别方法

4.1 对象在网格中的分布规律挖掘

在实体网格中一个重要的假设是“在连贯性的语篇中，实体的分布展现一定的规律性。”，在实体网格中通过计算实体间的转换模式来计算语篇的连贯性。而对象网格作为实体网格的扩展，仍然具有这样的假设，很多重要的规律隐士地被模式化在对象网格中。这种对象在网

格中的分布规律并不是随意的，它受到了向心理论[25]、齐普夫定律[26]以及其它基于实体的语篇理论的支持。另外，向心理论还认为对象被引入和提及取决于其在给定语篇中的全局角色。因此，本文将网格中的对象分为语篇显著对象和非显著对象，并将频率等于 1 的对象作为语篇非显著对象，在对象网格中将其去除，转换为由显著对象的构成对象网格。

为了把握重要对象在网格中的分布规律，用于对语篇重要对象的识别，笔者对大量不同类型的语篇对应的对象网格进行了分析。下面以美国能源部发布的新闻“Chu visits site of America’s first new nuclear reactor in three decades Department of Energy”[2]为例对重要对象分布的规律进行说明，其所对应的显著对象的对象网格如图 3 所示。

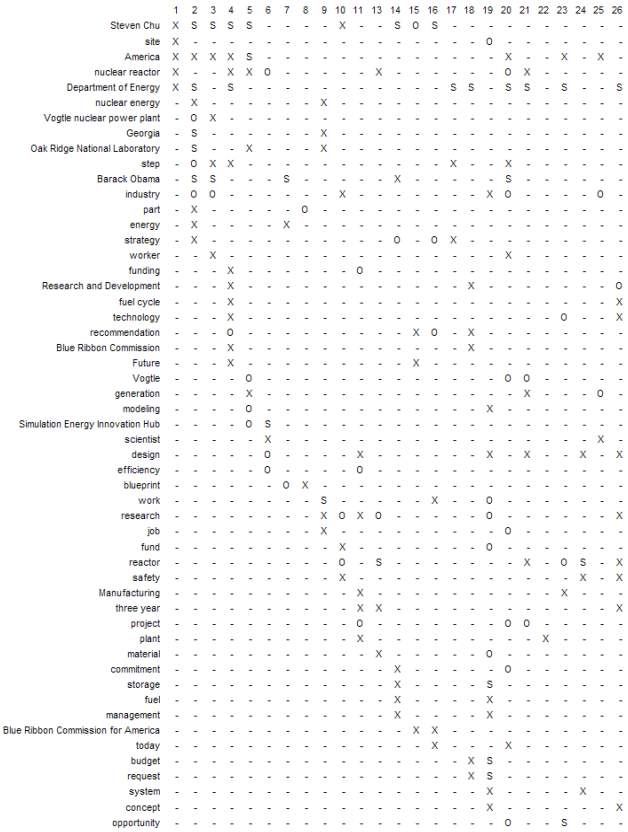

图 3 语篇显著对象的对象网格

向心理论根据中心的变迁，可以将整个语篇划分为一些相对内聚的片段，每个片段说明一个中心，即当前的主题，它或者是一个对象，通过不断的描写说明这个对象的某些状态或与这个对象相关的事件；它或者是某一事件，通过不断刻画说明该事件的条件、原因、结果、时间、地点等因素，以及与该事件相关的所有对象状态的变化。中心的变迁是指当前描述的主题发生了转变，新的问题成为关注的焦点[25]。据此，本文将在一个片段内频繁出现的对象作为论述该片段主题的一个对象，并对这个对象的分布进行平滑，并用“F”进行填充，例如：对象在相邻的四个句子中的分布模式为：“S – O X”，平滑过后为“S F O X”，图 4 为图 3 中对象网格经过平滑处理后其语法角色分布的状况。

[2] http://energy.gov/articles/chu-visits-site-america-s-first-new- nuclear-reactor-three-decades.

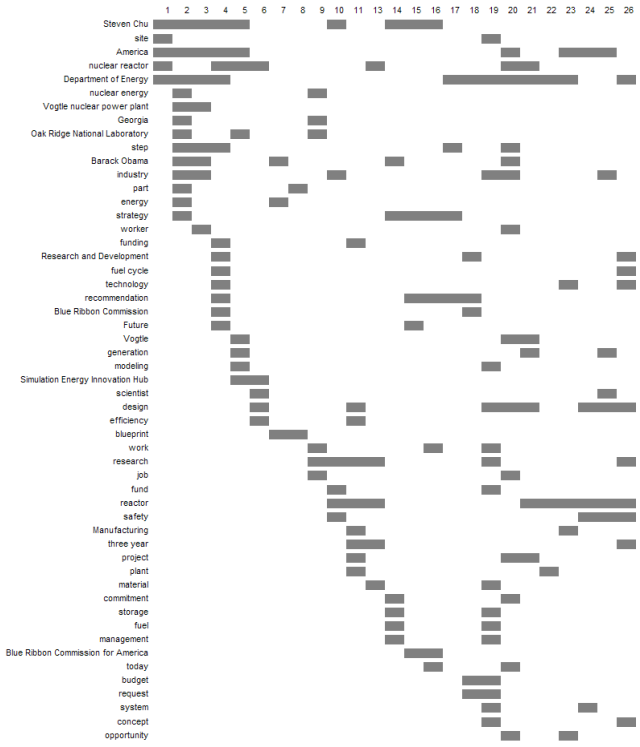

图 4 语篇显著对象在网格中的分布状况

从图 3 和图 4 中可以看出，重要对象在网格中的分布具有明显的规律性，具体为：

（1）在语篇中通常网格密度大的对象比密度小的对象更重要。所谓网格密度是指对象出现的网格频次和句子跨度的比例，代表了对象被集中讨论的程度，例如：对象网格中的“Steven Chu”、“reactor ”等。

（2）在语篇的对象网格分布中对象块（连续的非“-”分布）的长度越长，说明对象的“聚集”越高，对象往往越重要，例如：对象网格中的“America”、“research”等。

（3）重要的对象往往是命名实体、术语类型的对象，这些对象包含了丰富的语义含义，更能表达语篇的主题，例如：对象网格中的“Department of Energy”、“nuclear reactor”等。

（4）重要的对象往往出现在重要的语法角色上，比如：主语、宾语，例如：对象网格中的“Steven Chu”、“Department of Energy”等。

（5）对象在网格中的频度越高往往对象比较重要，例如：对象网格中的“America”、“Department of Energy”等。

（6）重要的对象往往出现在语篇的开始部分，例如：对象网格中的“Steven Chu”、“America”等。

（7）重要对象往往贯穿整个语篇，对象在语篇中被讨论的范围较大，例如：对象网格中的“nuclear reactor”、“Department of Energy”等。

但如何将这些规律转化为机器可读的可比较的定量指标，从中识别出网格中的重要对象，还需要做进一步算法上的设计。

4.2 网格分布规律的定量指标

基于网格分布的重要对象识别算法将以上重要对象在网格中的分布规律归纳为对象凝

聚度、对象活跃度和对象生命跨度三个指标，并分别进行计算。

➢ **对象凝聚度计算**

指同一个对象在语篇中的聚合程度，由对象块的长度、对象的密度以及对象自身的长度三个部分组成：

（1）对象块的长度，指对象所有块长度（所跨连续句子的数目）大于 1 的块的总长度。笔者提出块长度计算公式定义为公式 1。

$$blockScore(O,D) = log(\sum_{i=1}^{blockSize\ (O,D)} blockLength(O,i) + 10) \quad \text{公式 1}$$

其中，“*O*”表示对象，“*D*”表示语篇，$blockLength(O,i)$表示对象 *O* 第 *i* 个块的长度。

（2）对象的密度，指对象出现的网格频次和句子跨度的比例。笔者提出对象密度计算公式定义为公式 2。

$$density(O,D) = log(\frac{frequency(O,D)}{lastSentNum(O) - firstSentNum(O) + 1} + 10) \quad \text{公式 2}$$

其中，$frequency(O,D)$表示对象 *O* 在网格出现的频次，$firstSentNum(O)$指对象 *O* 第一次出现的句子序号，$lastSentNum(O)$表示对象最后一次出现的句子序号。

（3）对象的自身长度，指对象所包含的词的个数。笔者提出对象自身长度计算公式定义为公式 3。

$$lengthScore(O,D) = log(wordSize(O) + 10) \quad \text{公式 3}$$

对象的凝聚度的计算，是对以上三个指标进行综合，进行归一化处理，笔者提出对象凝聚度计算公式定义为公式 4。

$$cohesion(O,D) = \frac{lengthScore(O,D) * blockScore(O,D) * density(O,D)}{maxCohesion(D)} \quad \text{公式 4}$$

其中，$maxCohesion(D)$为语篇 *D* 中对象凝聚度的最大值。

➢ **对象活跃度计算**

对象活跃度指对象在语篇中的活跃程度，包括：对象的语法角色、对象的网格频度和对象首次出现位置三个指标。

（1）对象的语法角色。通常来说处于主语位置的对象要比处于宾语位置的对象活跃，而它们都比处于其它位置的对象活跃，本文为不同语法角色赋予的权值，如下所示：

$$roleWeight(role) = \begin{cases} 3 & role = S, \\ 2 & role = O, \\ 1 & role = X. \end{cases} \quad \text{公式 5}$$

（2）对象的网格频度，指的是对象在网格中出现的频次。它不同于语篇频次，因在网格构建过程中，同一对象多次出现在同一个句子中将会被合并，并取其中优先级最高的角色来表示对象的语法角色。在计算对象的网格频度时，同样采用语篇中网格频度最高的值进行归一化处理，笔者提出网格频度计算公式定义为公式 6。

$$frequencyRole(O,D) = \frac{\sum_{i=1}^{frequency\ (O,D)} roleWeight(role(O,i))}{maxFrequencyRole(D)} \quad \text{公式 6}$$

其中，$role(O,i)$表示对象 *O* 第 *i* 次出现时的语法角色，$maxFrequencyRole(D)$表示语

篇 D 当中对象网格频度最高的值。

（3）对象首次出现位置。一般来说首次出现位置越靠前对象越重要，比如出现在标题中的对象相对来说就很重要，笔者提出对象首次出现位置的计算公式定义为公式 7。

$$depth(O,D) = 1 - \frac{firstSentNum(O)}{sentSize(D)+1} \quad \text{公式 7}$$

对以上三个指标进行综合就构成了对象的活跃度，笔者提出对象活跃度的计算公式定义为公式 8。

$$active(O,D) = \frac{ln(depth(O,D) * frequencyRole(O,D) + e)}{maxActive(D)} \quad \text{公式 8}$$

其中，$maxActive(D)$为语篇 D 中对象网格频度的最大值。

- **对象生命跨度计算**

对象的生命跨度用于反映对象在语篇中的生命周期，指的是对象第一次被提及到最后一次被提及之间的距离，笔者提出对象生命跨度的计算公式定义为公式 9。

$$lifeSpan(O,D) = \frac{ln(\frac{lastSentNum(O) - firstSentNum(O) + 1}{sentSize(D)} + e)}{maxLifeSpan(D)} \quad \text{公式 9}$$

其中，$firstSentNum(O)$指的是对象 O 第一次出现的句子序号，而$lastSentNum(O)$表示其最后一次出现时的句子序号，$sentSize(D)$指语篇 D 中包含的句子数，$maxLifeSpan(D)$为语篇 D 中对象生命跨度的最大值。

4.3 对象重要度计算

对象重要程度的计算，需要将对象的凝聚度、对象活跃度、对象生命跨度三个指标进行综合，并进行归一化处理。笔者提出对象重要度的计算公式定义为公式 10。

$$objectWeight(O,D) = \frac{cohesion(O,D) * active(O,D) * lifeSpan(O,D)}{maxObjectWeight(D)} \quad \text{公式 10}$$

其中，maxObjectWeight(D)为语篇 D 中对象重要程度的最大值。

5 实验验证

为了验证基于对象网格的网络科技信息重要对象识别方法的有效性，本文将其识别的结果同 KEA 算法以及人工标注的结果进行对比分析。

5.1 实验数据

本实验选用 2011-09-01 到 2012-08-31 之间来自于美国能源部且被中国科学院武汉先进能源团队发布的《先进能源科技动态监测快报》选用的 30 条网络资源作为实验数据，其中 20 条资源作为测试数据，10 条资源作为 KEA 算法的训练集，对其进行采集和数据清理，并将网页主体内容保存为“TXT”文本格式。

5.2 实验过程

重要实验环境为：硬件：Intel 双核 2.94GHz CPU、4G 内存、64 位 Window 2008 操作系

统。开发环境：JDK1.6、MyEclipse 10 开发平台。开发语言：Java 语言。

主要开发工具：KEA-5.0、Tomcat6.0、stanford-corenlp-1.3.4.jar、stanford-tregex.jar、能源科技领域科研本体。

实验过程主要包括以下三个方面：

（1）构建基于对象网格的网络科技信息重要对象识别原型系统。笔者利用 stanford-corenlp-1.3.4.jar、stanford-tregex.jar 和能源科技领域科研本体设计实现了知识对象抽取、知识对象的共指识别、语义相关对象识别以及对象的语法角色识别（限于篇幅这些算法的实现过程不在累述），并最终由这些算法构成了重要对象识别原形系统，通过该系统能够实现网络科技信息从输入到输出重要对象及其重要程度的整个过程。

（2）使用人工标注语篇的重要对象作为对比的基准数据。对于网络科技信息来说，一般不包含关键词或标签，因此，笔者请三个情报专业的研究生对每篇语篇标注 6-10 个重要对象，并请能源领域的战略专家对标注后的结果进行审核，确定最终的标注结果，作为验证的基准数据。

（3）选用关键词抽取工具 KEA-5.0（Keyphrase extraction algorithm）作为本文提出方法的对比对象。KEA 是 Witten 等开发的一套关键词抽取工具[7]，它能够对“SKOS”以及“TXT”格式的词表进行支持，还支持从训练集中提取训练模型来提高关键词识别的准确度。为了让 KEA 算法的性能达到最大，本文将能源领域的科研本体转为化 SKOS 格式的词表，用于对 KEA 工具的支持，并将人工标注的 10 篇语篇作为 KEA 算法提取训练模型的训练集。

5.3 实验结果与分析

将本文提出算法的识别结果同人工标注和 KEA 算法识别的结果进行精确匹配和近似匹配。精确匹配直接和基准数据进行一对一的匹配。近似匹配基于语义相关度，认为与基准数据中的对象词义高度相近的对象也可以被看作是揭示主题的重要对象。例如，人工标注重要对象“Advanced Research Projects Agency-Energy Programs”，如重要对象识别算法的结果为“Advanced Research Projects Agency-Energy”，也看作是匹配成功。在本实验中只有一个对象完全包含另一对象才会被认为近似匹配成功。

本文提出的重要对象识别方法和 KEA 算法对语篇中的所有对象进行识别并按照其重要程度进行排序，但人工对测试语篇标注的重要对象数量平均为 8.1 个，为了在准确率和召回率之间进行平衡，实验中分别选取本文提出的方法和 KEA 算法识别结果的前 9 个重要对象进行对比，其对比结果如表 1 所示。

表 1 基于人工标注重要对象的各方法计算结果对比

基准数据	人工标注的重要对象			
指标	准确率		召回率	
匹配模式 / 识别方法	精确匹配	近似匹配	精确匹配	近似匹配
KEA 算法	40.56%	51.11%	47.32%	59.46%
本文提出的方法	45.00%	54.44%	52.15%	63.32%

在准确率方面，各方法的准确率均不是很高，经过分析发现主要是由两个原因造成：（1）因为人工标注的重要对象只有在科研本体中时，才会对这些对象进行规范，例如：“DOE”

会被规范为“Department of Energy”，但有些对象，例如“building efficiency”并不在科研本体当中，对这些术语类型的对象的自动识别，往往需要一定数量训练集，对于本文提出的方法在进行重要对象识别时，并没有训练集的辅助。所以，对于此类对象并不能进行很好的识别，这也是造成准确率不高的主要原因；（2）由于人工标注的重要对象平均为 8.1 个，而各个方法选用的重要对象为 9 个，分母相对偏大也是导致准确率不高的一个原因。

虽然精确匹配上，各方法的百分比不高，但仍可以看出本文提出方法的准确率高于 KEA 算法。从表 1 中可以看出各方法的近似匹配效果均优于精确匹配效果，通过近似匹配，弥补了精确匹配度较低的不足，拓展了结果中描述语篇主题的相似对象，在实际语篇重要对象识别中这种匹配更有意义。通过对比结果可以看出，本文提出的方法比 KEA 在近似匹配上更优势，准确率提升到 54.44%。在召回率方面，影响其大小因素和影响准确率的第一个因素类似，但由于人工标注的平均为 8.1 个，小于测试中选取的 9 个，所以，召回率要高于准确率。通过对比结果可以看出，本文提出的方法比 KEA 算法在准确率和召回率上都具有优势，另外，与 KEA 算法相比，在识别过程中本文提出的方法并没有使用人工标注的训练集，而且 KEA 算法在基于词表进行的知识对象抽取中还存在一定的误判情况，进一步验证了本文提出方法的优越性。

6 总结

为了对网络科技信息中包含的重要情报线索进行揭示，需要一个能够蕴含更多对象特征以及语篇结构信息的表示模型，将非结构化的语篇映射成可以计算的知识单元。通过对语篇表示模型包含信息的挖掘和计算实现对语篇重要对象的识别。本文围绕以上问题展开研究，提出基于对象网格的语篇表示模型，利用对象在网格中的分布规律对语篇中的重要对象进行识别。通过将本文提出的方法同 KEA 算法和人工标注的结果进行对比，验证了基于对象网格的网络科技信息重要对象识别方法的有效性。在未来的工作中将进一步提高算法的准确度，并验证其在不同领域网络科技信息中的效能。

参考文献

[1] A strategy for American innovation: Securing our economic growth and prosperity[EB/OL]. [2013-01-09]. http://www.whitehouse.gov/sites/default/files/uploads/InnovationStrategy.pdf.

[2] Obama administration unveils "big data" initiative: Announces $200 Million in new R&D investments[EB/OL]. [2013-01-09]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[3] Innovation Union Scoreboard[EB/OL]. [2013-01-09]. http://ec.europa.eu/enterprise/policies/innovation/facts-figures-analysis/innovation-scoreboard/index_en.htm.

[4] Main Science and Technology Indicators (MSTI): 2012/1 edition[EB/OL]. [2013-01-09]. http://www.oecd.org/science/innovationinsciencetechnologyandindustry/mainscienceandtechnologyindicatorsmsti20122edition.htm.

[5] 邹益民, 张智雄, 刘建华. 基于对象行为的情报关注模型研究[J]. 中国图书馆学报, 2013, 39(207): 50-59.

[6] Frank E, Paynter G W, Witten I H, et al. Domain-specific keyphrase extraction[C]. Proceedings of the 16th International Joint Conference on Artificial Intelligence(IJCAI-99), 1999: 668-673.

[7] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]. Proceedings of the 4th ACM Conference on Digital Libraries(Digital Libraries'99), Berkeley,CA, USA, 1999: 254-255.

[8] Turney P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2000, 2(4): 303-336.

[9] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[C]. Proceedings of the 7th Web-Age Information Management(WAIM'06), Hong Kong, China, 2006: 85-96.

[10] Cohen J D. Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting[J]. JASIS, 1995, 46(3): 162-174.

[11] Luhn H P. A statistical approach to mechanized encoding and searching of literary information[J]. IBM Journal of research and development, 1957, 1(4): 309-317.

[12] Salton G, Yang C-S, Yu C T. A theory of term importance in automatic text analysis[J]. Journal of the American society for Information Science, 1975, 26(1): 33-44.

[13] You W, Fontaine D, Barthès J-P. An automatic keyphrase extraction system for scientific documents[J]. Knowledge and Information Systems, 2013, 34(3): 1-34.

[14] Kim S N, Kan M-Y. Re-examining automatic keyphrase extraction approaches in scientific articles[C]. Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications(MWE'09), Singapore, 2009: 9-16.

[15] Ferrara F, Pudota N, Tasso C. A keyphrase-based paper recommender system[C]. Proceedings of the 7th Italian Research Conference on Digital Libraries(IRCDL'11), Pisa, Italy, 2011: 14-25.

[16] Kastner I, Monz C. Automatic single-document key fact extraction from newswire articles[C]. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics(EACL'09), Athens, Greece, 2009: 415-423.

[17] Rahma A M S, Kadhem S M, Farhan A K. Finding the Relevance Degree between an English Text and its Title[J]. Eng. & Tech. Journal, 2012, 30(9): 1625-1640.

[18] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing(EMNLP'04), Barcelona, Spain, 2004.

[19] Litvak M, Last M. Graph-based keyword extraction for single-document summarization[C]. Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization(MMIES'08), Manchester, UK, 2008: 17-24.

[20] Huang C, Tian Y, Zhou Z, et al. Keyphrase extraction using semantic networks structure analysis[C]. Proceedings of the 6th IEEE International Conference on Data Mining(ICDM'06), Hongkong, China, 2006: 275-284.

[21] Barzilay R, Lapata M. Modeling local coherence: An entity-based approach[C]. Proceedings of the 43rd Annual Meeting of the ACL(ACL'05), Ann Arbor, USA, 2005: 141-148.

[22] Barzilay R, Lapata M. Modeling local coherence: An entity-based approach[J]. Computational Linguistics, 2008, 34(1): 1-34.

[23] Lapata M, Barzilay R. Automatic evaluation of text coherence: Models and representations[C]. Proceedings of the19th International Joint Conference on Artificial Intelligence(IJCAI'05), Edinburgh, Scotland, UK, 2005.

[24] 邹益民，张智雄，曲云鹏. 基于实体网格的语篇表示模型研究[J]. 情报理论与实践，

2013, 36(6): 102-106.
[25] Walker M A, Joshi A a K, Prince E E F. Centering theory in discourse[M]. Oxford: Oxford University Press, 1998.
[26] Newman M E. Power laws, Pareto distributions and Zipf's law[J]. Contemporary Physics, 2005, 46(5): 323-351.
[27] KEA description[EB/OL]. [2013-04-11]. http://www.nzdl.org/Kea/description.html.