基于知网的相关概念场的构建

董 强　　董振东
中国科学院计算机语言信息工程研究中心 北京 100083
DongQinag@keenage.com　　dzd@keenage.com

摘 要：词语的相关性及其知识的获取是人类语言技术研究中的热点之一。国内外关于这个方面以及与之相关的研究已有不少的报道。 我们利用知网知识系统在这方面做了有意义的研究，提出了相关概念场的概念，构建了知网相关性激发器，取得了令人满意的结果。本文讨论了词语相关性与词语相似性的不同和词语相关性与概念相关性的不同。本文阐明了相关概念场的理念，介绍了它的实现的原理、方法，以及它的潜在的应用。

关键字：词语相关性；相关概念；相关概念场；相关性激发器；知网

Construction of HowNet-based Relevant Concept Field

Qiang Dong　　Zhendong Dong
Research Center of Computer Language Engineering, Chinese Academy of Sciences, Beijing, 100083
DongQinag@keenage.com　dzd@keenage.com

Abstract: The acquisition of the knowledge of word relevancy is one of the hot topics in human language technology. A lot of reports on this subject and relevant studies have been published. We have applied HowNet knowledge system to the subject these years, proposed Relevant Concept Field, and successfully constructed a Relevancy Activator. This paper discusses the difference between word relevancy and word similarity and the difference between word relevancy and concept relevancy. The paper describes the conception of Relevant Concept Field, presents the principle and method for its visualization, and proposes its potential applications.

Keywords: Word Relevancy; Relevant Concept; Relevant Concept Field; Relevancy Activator; HowNet

1、问题的提出

由于因特网的出现，人们对于信息处理的需求的日益提高，词语间关系的研究的广度和深度也在不断的加强。词语间关系的研究在信息处理的多个方面都有着广泛的应用，例如语义排歧、文本的自动聚类、分类、信息检索等等。近两年我们在完善和扩充知网的建设的同时，也对词语间关系进行了深入的研究。我们的研究主要集中在两方面，一是词语的相关性；一是词语的相似度。本文介绍的是我们的关于词语相关性的研究。这一研究的核心是：构建基

于知网的相关概念场。我们想先说明两个问题：第一，词语相关性与词语相似性的不同；第二，词语相关性与概念相关性的不同。

1.1 词语相关性与词语相似性的不同

词语相关性有时容易与词语相似性相混淆，其实它们是两个完全不同的概念。词语相关性反映的是两个词语互相关联的程度，即词语之间的组合特点，它可以用词语在同一个语境中共现的可能性来衡量。而词语相似性反映的是词语之间的聚合特点。刘群等曾利用知网进行了词语相似性的研究[2]。相关概念场则是利用知网进行了词语相关性的研究。通过比较这两个软件，可以比较出两个概念的差别。用刘群的软件对“旅游”、“名胜”、“导游”、“安检”、“游客”、“学校”、“学生”进行相似度测算，将得到如下数据：

“旅游”和“游客”：0.166698　　“向导”和“游客”：0.891667
“导游”和“安检”：0.044444　　“向导”和“学生”：0.889623
“游客”和“学校”：0.121061　　“宾馆”和“学校”：0.578654

上述数据在人的直觉上是合乎道理的。但采用《知网相关概念场》软件对“旅游”、“安检”、“宾馆”、“学校”进行相关性测算，其具体数据则是：

“旅游”：274 词语，其中包括“安检”和“宾馆”，但不包括“学校”、“学生”
“安检”：148 词语，其中包括“旅游”和“宾馆”，但不包括“学校”、“学生”
“宾馆”：1524 词语，其中包括“旅游”和“安检”，但不包括“学校”、“学生”
“学校”：546 词语，其中包括“学生”，但不包括“旅游”、“安检”、“宾馆”
“学生”：545 词语，其中包括“学校”，但不包括“旅游”、“安检”、“宾馆”

观察测试的结果我们会发现“宾馆”和“学校”的相似度不低，达到 0.578654，但是在它们的相关概念场中，“宾馆”的相关概念场中并不包括“学校”，“学校”的相关概念场中也不包括“宾馆”。因此这两个概念之间是不同的。

1.2 词语相关性与概念相关性的不同

词语相关性与概念相关性的不同表现在两个方面：第一，词语相关性指的是词语的形式，而概念相关性指的是词语的意义；第二，词语相关性虽然基本上基于自身的意义，但它们有可能是与具体语言相关的，它们包含因语言习惯而引起的相关，例如，中文里“首脑”和“领导人”，在概念上并没有什么差别，应是属于一个同义集，但是后者与“国家”的共现率比与前者的高得多（在 Google 上查“国家领导人”出现次数是“国家首脑”的 10 倍）。但是从概念的相关性角度，“领导人”似乎可以与任何表示“机构”这一概念相关。

目前国内外都有不少学者致力于词语间关系的研究，提出了一些有意义的方法。这些方法不外乎两种，一是手工的，例如现有的各种义类词典、搭配词典等；另一是基于大规模的语料库进行自动获取或称无导获取[3]。但是，自然语言本身的特性决定了词语组合之间的不定性，有很多种组合并不是常见的，甚至有很多的组合情况在语料库中是没有的，因此这种方法往

往受制于所采用的语料库，难以避免数据稀疏的不足。
而且迄今为止，对于词语的相关性的研究往往是基于词语的词形，而不是基于词语所表达的概念。例如：英文词语“bank”，它有两个最常见的意义，一是“银行”；二是“河岸”。如果是基于词语词形的研究，则在得到的结果中既包含所有与“银行”相关的词语，也包含与“河岸”相关的词语，可见在得到的结果中含有的噪声是相当大的。另外，一般对于词语相关性的研究都是小范围的，单语种的。

2、相关概念场

2.1 什么是相关概念场

这里所说的相关是指不同的概念在某种语境中共现的可能性。相关概念是指词语所代表的概念与哪些概念相关。相关概念场是一个相关概念的集合，是与一个词语的某个概念相关的所有概念的集合。相关概念场中的相关概念自然是由词语来表现的，但这些词语所代表的概念已经是唯一的，而不再是有歧义的。在理论上，同一个词语由于其概念的差异将有若干个不同的相关概念场，一个词语有多少个概念就应该有多少不同的相关概念场。这点也已为我们的实验所证明。相关概念场源于不同的意义，同时又框定了特定的意义。我们的相关概念场是指某一特定的词语所代表的特定概念被激发而形成的相关概念的集合，我们把我们所开发的包含所有相关概念场的数据库称为相关概念场库。

2.2 基于知网的相关概念场的特点

我们利用知网知识系统进行了基于概念的、大规模、双语种的词语相关性研究，并成功地开发了相关概念激发器，构建了基于知网的相关概念场库。基于知网的相关概念场有如下特点：

a. 基于概念。由于我们认为词语的相关性最终应归结于词语所代表的概念的相关性，因此我们的研究方法是基于概念的方法。
b. 大范围。基于知网的相关概念场涵盖了知网知识库中所有的词语及其所代表的义项。目前为中英文词语各 6 万 5 千多，概念各 8 万多。
c. 跨语种（cross-language）。知网知识系统是一个中文与英文双语的系统，相关概念场也是双语的。由于我们是基于概念的，激发器的基础是义原和关系，因此它是独立于特定语言的。

例如查找词语“营”，这个词语在知网知识库中共有 4 个意义，分别是：

DEF={facilities|设施}
DEF={manage|管理}
DEF={part|部件:domain={military|军},whole={army|军队}}
DEF={seek|谋取}

当我们选定其中的第三个概念，即 DEF={part|部件:domain={military|军},whole={army|军队}}，

表示“军营”的意思。通过查找相关概念场就会得到1348个词语，如：战况、军事学院、战斗力、兵团等等，也就是说，与“营”这个词语相关的词语共有1348个，因此从表向上看，它们是词语与词语之间的关系。但是当我们在这1348个词语中选定一个词语，如“连”的时候，我们会发现“连”的定义是：DEF={part|部件:domain={military|军},whole={army|军队}}，也就是说，“连”在作为“连队”这个意义的时候与“营”作为“军营”讲的时候是相关的。而“连”还有其他的意义，如表示“连接”，但是它在作为“连接”这个意义时与“营”作为“军营”讲的时候并不相关。由此可见，我们的相关概念场，是基于词语的概念的。

对于英文也是类似的，例如查找英文词语“river”，在作为“江、河”讲的时候会得到222个相关的英文词语，如：flow（流量），flood （发大水），riverbed （河床）等。我们将一些相关概念场得到的结果放在附录中供大家参考。

2.3 基于知网的相关概念场的实现

2.3.1 实现的理念

知网知识系统的特点决定了相关概念场实现的可能性，关键点在于：

a. 知网知识系统是描述概念与概念，概念的属性与属性之间的关系的知识系统。知网知识系统在设计的目的上就是为了可以体现及运算概念之间的关系，即相关概念场所体现的正是设计者的目的。

b. 知网知识系统中共有义原（event|事件，entity|实体，attribute|属性，AttributeValue|属性值等）2200个左右，关系（agent, patient, target, time, etc.）110个。

c. 知网知识库的描述体系是一套便于结构化表达的描述语言。

正是上述的特点决定了知网知识系统便于高效率、高准确度的意义计算。然而，知网知识库中对每一个概念的描述是孤立的、静止的，那么概念之间的相关性又是如何被实现的呢？下面我们先来观察“大学”和“大学生”这两个词语，从中可以看到我们的做法。

DEF={InstitutePlace|场所:domain={education|教育},modifier={HighRank|高等},{study|学习:location={~}},{teach|教:location={~}}}

这里InstitutePlace|场所、education|教育、HighRank|高等、study|学习、teach|教等均为义原。而domain、location、:modifier等均为关系。这一段知识描述语言所表达的意义是直观的，即是“大学”是一个场所，属于教育领域，等级为高等的，在这个场所里人们学习和教书”。

DEF={human|人:{study|学习:agent={~},location={InstitutePlace|场所:domain={education|教育},modifier={HighRank|高等},{study|学习:location={~}},{teach|教:location={~}}}}}

这一段知识描述语言所表达的意义也是直观的，即“大学生”是一个人，他是学习的施事，他在上面描述的场所学习。

从这两个词语的定义中可见我们对于这两个词语的描述是静止的和孤立的，我们并没有描述它们之间有什么关联。但大家一定可以从它们各自的定义中看到它们是通过某些义原而产生特定的关系。

2.3.2 实现所采用的资源

基于知网的相关概念场在实现时采用了以下知网所提供的资源：

a. 知网知识库　　b. 知网事件类义原表

c. 知网实体类义原表　　d. 知网属性类义原表

e. 知网属性值类义原表　　f. 知网事件角色与关系

g. 知网反义表　　h.知网对义表

2.3.3 实现的方法

在知网知识库中我们对概念的描述是孤立的、静止的，为了实现概念之间的动态联系，我们构造了一套关系激活机制，我们称之为相关性激发器。在这套机制中包括 DEF 分析器、规则匹配器、构造器等部分，其中规则分类器为这套机制的核心部分，它决定了最后输出的结果。我们首先通过 DEF 分析器把不同的概念描述划分为简单概念和复杂概念，再进一步利用知网义原表将其分为四大类：实体类、事件类、属性类和属性值类。如果是复杂概念，需要在前一步的基础上提取出义原之间的各种关系。我们对各种类型的关系编写了相应的规则，以便于根据不同的关系提取与之相关的概念。规则匹配器就是为了完成这一工作所设计的，它根据提取出的各种关系规则匹配，从而获得所有与之相关的概念。我们通过构造器对所得到的结果进行去重、排序和整理。最后，输出结果。

3、相关概念场的应用

3.1 相关概念场在排歧中的作用

相关概念场为语义排歧提供了一种新的手段。在文本分析时，如果遇到一个需要排歧的词语，那么首先利用知网知识库，得到这个词语的所有概念。接下来，利用知网相关概念场得到每一种概念的相关概念场，那么这个词语的有多少个概念就会形成相应个数的概念场。最后将该词语所在语境中的其他词语，分别投射到这几个场中，通过比较这几个场密度的大小，从而可以得到该词语的意义。

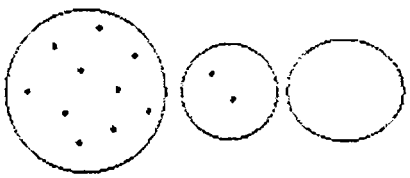

图 2：相关概念场在排歧中的作用

3.2 相关概念场在文本自动聚类中的作用

一段文字中所有词语的相关概念场的集合构成这段文字的相关概念场，我们可以根据两个场之间相重合的程度判断这两段文字之间的关系。如果两段文字的场之间的重合度高，则可以

认为它们是一类文本，反之，则分属于不同的类别。如图所示，A 为文本 A 生成的概念场，B 为文本 B 生成的概念场，如果 A 场包含 B 场，则文本 A 和 B 为同一类。如果，A 与 B 部分重合，则需要计算重合的系数，根据系数的大小来判断 A 与 B 是否为同一类。如果 A 与 B 完全不重合，则 A 与 B 属于不同的类别。

图 3：相关概念场在自动聚类中的作用

4、结论

知网知识系统将概念分解为具有唯一的确定意义的义原，并通过各种不同的关系将这些义原组织起来描述一个概念，它的这种结构化的描述语言为意义的计算提供了方便。知网相关概念场以知网知识系统为基础，充分利用知网知识系统中的各种资源，得到的结果与人的直觉比较符合。与传统方法相比，它脱离了语料的束缚，仅仅利用一个资源就实现了大规模的概念之间语义关系的计算，并且它是跨语种的。通过大量的实验观察，取得的结果令人十分满意。并且它的成功为自然语言处理中的其他领域提出了一种新的解决办法，具有很高的实用价值。它的成功也进一步证实了知网知识系统的在意义计算方面的能力，也说明知网知识系统是一个具有实际使用价值的语义资源。

参 考 文 献

[1] Dagan I., Lee L. and Pereira F. (1999), Similarity-based models of word cooccurrence probabilities, Machine Learning, Special issue on Machine Learning and Natural Language, 1999

[2] 刘群，李素建，基于《知网》的词汇语义相似度计算，第三届中文词汇语义研讨会论文集

[3] 鲁松，白硕，(2001) 自然语言处理中词相关性知识无导获取和均衡分类器构建，中科院计算所，博士论文

[4] 董振东，董强 (1999)，“知网”，http://www.keenage.com

[5] 杨晓峰，李堂秋 一种基于知网的语义排歧模型研究，中文计算语言学期刊，Vol.7，No.1，2000, pp47-78

[6] 李涓子，汉语词义排歧方法研究，清华大学博士论文，1999

附　　录

“吃”（各义项）

DEF={absorb|吸收}　　相关概念 13 个

吸湿性;有吸收力;有渗透性;吃;摄;摄取;吸;吸取;吸收;吸热;吸湿;养分;养料;

DEF={depend|依靠}　　相关概念 76 个，因篇幅关系取前 50 个

超脱；出世；独立；独立自主；各奔前程；各行其是；各自为政；貌合神离；自给自足；自立；自力；自力更生；自食其力；相互依存；有条件；季节性；依赖性；吃；负；附；寄；据；看；靠；赖；赖以；凭，凭借；身不由己；恃；托；托福；相互依赖；相依；相依为命；仰；仰承；仰人鼻息；仰仗；依；依傍；依附；依靠；依赖；依凭；依托；依仗；倚；倚仗；倚重；

DEF={destroy|消灭}　　相关概念 966 个，因篇幅关系取前 50 个

杀伤力；杀伤性；弹药携行量；敌占；哀兵必胜；荷枪实弹；军情；敌情；阵势；阵；阵列；阵容；疑阵；后坐；退兵；退却；军备竞赛；兵工厂；兵种；增援；精兵简政；撤兵；撤防；撤军；镆铘；枝；支；发；门；梭子；尊；番号；赤手空拳；制空权；制海权；兵权；军权；手无寸铁；徒手；养兵；军力水平；兵力比；行伍出身；劳师；行军；急行军；射速；军籍；

DEF={eat|吃}　　相关概念 2011 个，因篇幅关系取前 50 个

饭量；食量；可食用性；口；口齿；草菇；冬菇；菇；海带；口蘑；蘑；蘑菇；松蕈；单产；存栏；极量；剂量；剂量大小；药量；用量；香；口福；魔芋；牧草；好吃；贪吃；贪嘴；餐；斋戒；下马；食草；食肉；食性；宾馆；店；饭店；国宾馆；假日酒店；酒店；客栈；旅店；旅馆；旅舍；旅社；迎宾馆；栈房；招待所；果园；苹果园；葡萄园；

DEF={exhaust|损耗}　　相关概念 33 个

耗电量；油耗；用量；用水量；吃；耗尽；耗散；耗损；耗用；竭；竭尽；尽；损；损耗；拖垮；消费；消耗；消磨；消损；用尽；讫；殚；铄；罄；耗散；耗能；耗时；吃水；低耗；消费税；内耗；内耗；能耗；

DEF={suffer|遭受}　　相关概念 265 个，因篇幅关系取前 50 个

抵抗力；免疫功能；免疫力；砸饭碗；绳之以法；兵连祸结；樊笼；自讨苦吃；免疫；屈打成招；保命；大难不死；虎口余生；劫后余生；死里逃生；逃过一劫；幸免于难；人不犯我，我不犯人；人若犯我，我必犯人；吃亏长见识；饮弹身亡；饱；吃香的喝辣的；赏；赏玩；受用；松快；体味；玩；玩儿；玩赏；享；享受；享用；消受；欣赏；免于起诉；免罪；冷板凳；冷眼；冷遇；冤案；冤假错案；冤情；冤狱；连体孪生子；连体双胞胎；患难夫妻；小媳妇；灾民；

“先生”（各义项）

DEF={human|人:HostOf={Occupation|职位},domain={education|教育},{teach|教:agent={~}}}

相关概念 255 个，因篇幅关系取前 50 个

校风；同等学历；学历；本科；大本；大学程度；科教兴国；返校；绕场；校际；教会学校；北大；北京大学；复旦；复旦大学；南大；南京大学；南开；南开大学；清华；清华大学；同济大学；西安交大；中山大学；国防大学；校办工厂；特拉维夫大学；朱拉隆功大学；剑桥；剑桥大学；牛津；牛津大学；联合国大学；哈佛；哈佛大学；卡内基梅隆大学；麻省理工学院；耶鲁大学；西点；西点军校；党校；村塾；村学；体育学院；农大；农学院；美术学院；美院；医大；医科大学；

DEF={human|人:HostOf={Occupation|职位},domain={medical|医},{doctor|医治:agent={~}}}

相关概念 589 个，因篇幅关系取前 50 个

抵抗力；免疫功能；免疫力；结痂；医德；医风；脊髓麻醉；局部麻醉；局麻；麻醉；全麻；全身麻醉；针刺麻醉；活血；病况；病情；病势；病态；伤情；伤势；疫情；疗效；辍；医大；医科大学；医学院；教学医院；实习医院；病院；合同医院；救护所；门诊部；门诊所；卫生所；卫生院；医疗站；医务所；医院；院；诊所；中医医院；中医院；精神病院；传染病院；急救站；急救中心；疗养院；休养所；军医院；医疗后送站；

DEF={human|人:modifier={male|男}}相关概念 362 个，因篇幅关系取前 50 个

男子气；入赘；赘；重男轻女；夫权；男礼服；男装；望子成龙；男篮；男足；男排；男单；男双；鹊桥相会；生；小生；厮；舞男；他；表兄弟姐妹；表兄妹；俦；阿弟；阿哥；胞弟；胞兄；表弟；表哥；表兄；表兄弟；长兄；大哥；大舅子；弟；弟弟；哥；哥哥；姑嫂；姐；姐夫；姐姐；姐丈；襟；舅；舅子；老兄；妹；妹夫；妹妹；妹子；