个性化知识的表示方法

刘冬明[1]，杨尔弘[2]

（1.中北大学，山西省太原市 030051；2. 北京语言大学，北京市 100083）

摘要：在当前信息暴涨的时代，网络信息正在面临着各取所需的要求，信息检索、话题检测、信息推荐等应用技术都逐渐开始面向个性化的发展趋势。然而目前的个性化技术大都依赖于对用户行为的了解，根据用户的历史行为，判断和预测用户的目的，没有同用户的当前所具有的知识结合起来。本文提出一种用户个性化知识的粗略表示方法——词形关系图，作为个性化应用技术的基础。具体内容包括：词形关系图表示知识的方式，结合遗忘规律从用户语料库中获取个性化词形关联的方法，以及结合实验结果对这种表示方法应用可行性的初步分析。

关键词：知识表示；个性化；词形关系图

中图分类号：TP391　　**文献标识码**：A

The Representation method of Personalized Knowledge

LIU Dongming[1], YANG Erhong[2]

(1.North University of China ,Taiyuan,Shanxi　030051, China ; 2.Beijing Language and Culture University, Beijing　100083 ,China)

Abstract: Now, along with the sudden increase of the amount of network information, the personalization is becoming the critical component of information retrieval, topic detection, and information recommendation. However, most of the current personalization technologies rely on user behavior, which is based on the historical actions of the user to determine and predict the user's destination, not combining with a user's knowledge. This paper presents a method of rough knowledge representation for individual users - word relationship graph as the basis for personalized applications. Topics include: the method of knowledge Representation using word relationship graph, the acquirement of word relationship graph combined the forget law from the user corpus, and the experimental results analysis for the preliminary feasibility of this representation.

Key words: Knowledge Representation; Personalization; Word Relationship Graph

1 引言

在当前的信息处理中，针对海量的数据，快速准确的为用户提供合适数据的关键之处在于：一是数据的压缩，由于相同的信息在网络中存在大量重复的内容，如何精简压缩并以友好的形式呈现成为了获取信息之后关键要处理的问题，目前多文档文摘[1]、搜索结果多样化[2]等研究的目标就在于此；二是提供给用户的信息要同用户的需求以及用户的所具有的知识结合，由于用户往往难以准确描述自己想要的信息，这样系统必须对用户有所了解，才能真正提供用户所要的信息，当前许多这样的系统，推荐系统如购物网站通过顾客的浏览或购买历史给顾客推荐产品[3]，社交网络通过顾客的历史记录推荐感兴趣的人[4]，搜索系统如google 同 twitter 的合作，bing 同 facebook 的合作就源于此。

本文重点关注第二个问题。当前的解决方案中，系统对用户的了解或认识来自于用户的在某些网站的行为，即根据用户过去的行为判断用户当前的意图，这种解决方案有两个局限性,其一是信息来源受限,事实上任何一个网站或应用程序无法获知全部的用户浏览信息,技术上和法律上都存在着障碍;其二是用户的某些浏览未必就会在用户脑海中留下印象，经常

作者简介：刘冬明（1972——），男，讲师，自然语言处理；杨尔弘（1965——），女，教授，自然语言处理，计算语言学。

发生的状况是用户点击某些链接之后马上关闭，或者在社交网站中，用户也许追踪了一些人，但是他仅仅关心这些人发布的某些内容。

本文认为最佳的解决方案在于根据用户所具备的知识来解决此问题。假设每个用户所具备的知识能够形式化的表示出来，并且易于被应用程序所使用，那么这种个性化的知识同应用相结合就能够切实从海量的数据中挖掘出用户真正的需求。例如，同样是搜索“苹果”，一个信息技术爱好者会关注苹果公司的新技术新产品，而农业专家关心的苹果的培育技术，即使同指的是吃的“苹果”，食品安全行业的人、水果消费者、果树种植者等想获得的信息也不一样。又如当某一案件发生的时候，需要对这一话题追踪，普通民众关心受害者、嫌疑犯等，法律工作者可能更加关注及到的立法、司法等方面的状况。可见对于相同的用户表述形式，具有不同背景知识的用户需求不同。如果相关应用能够获取用户当前所具备的知识并且有效的利用用户的知识层面，那么问题将容易获得解决。因此问题的关键在于用户的知识如何有效的表示和使用，本文对此作了初步的探讨。

下面第二节介绍相关研究，第三节详细描述用户知识的粗略表示方案，第四节描述这种表示的获取方式，第五节进行实验结果分析，最后是总结和展望。

2 相关研究

本文致力于针对不同用户构建个性化的知识库，直接应用于海量数据的检索、分析并给予用户个性化的反馈。由于目前并没有直接相似的研究，下面仅仅给出知识库和个性化搜索或推荐的相关研究。

支持自动的文本分析以及人工智能应用的自然语言知识库有国外著名的 WordNet[5]和国内的 HowNet[6]等，它们的目的在于使常识性的知识能够计算化，在这些资源的基础上，近年来国外启动了多个著名的大规模的事件语义资源开发项目，如 FrameNet[7]、OntoNotes[8]等，从不同角度对英语真实文本句子中的事件语义信息进行了深度标注。而国内也针对汉语的研究现状，结合汉语自身的特点，设计并实现了“一个针对汉语客观事件的句法、语义和概念描述知识库——汉语事件知识库”[9]。这一类知识库不论是才有标注还是构建，都有一定深度，能够满足语义的深层次计算，但是对于本文所针对的问题来讲过于复杂以至于难以应用，同时它们基于人的普遍共识而不是个性化的建模。

用户与资源间的相关性、资源间的相似性计算是当前个性化搜索、兴趣推荐等研究关注的核心，而查询扩展、用户反馈、协同过滤等技术是这一领域的经典技术，对于用户的知识的建模仅仅局限于用户的行为，如采用流行的 folksonomy[10]等构建用户模型，这种研究方法本质在通过用户的外在的历史行为预测用户的未来行为，缺失了用户自身的知识库结构，结果必定受限。

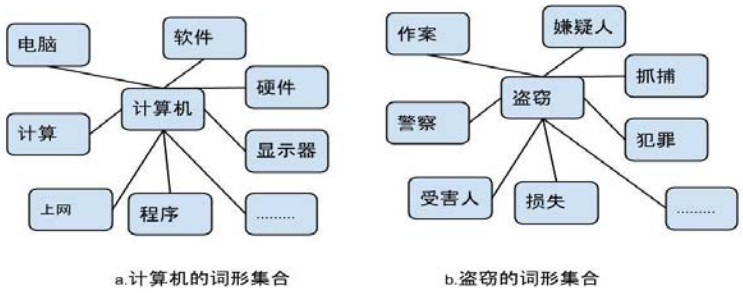
3 个性化知识的粗略表示

建立知识的完备表示方案对于现在的技术来讲是不现实的，本文仅仅针对基于个性的信息搜索、话题的检测跟踪、兴趣追踪等应用提出一种粗略的表示方法。

根据知网对知识的定义:“知识是一个系统，是一个包含着各种概念与概念之间的关系，以及概念的属性与属性之间的关系的系统”。知网的目标在于对常识的计算，因此定义了义原、属性、各类关系等形式来描述和表示知识。本文基于前面提到的应用目的认为自然语言是对知识天然的描述工具，词作为最小的语义单位，通过词之间的关系能够对知识进行粗略的表示，即能够粗略的表示出概念的属性与属性之间的关系以及概念之间的关系。一个人同另外一个人知识的不同即在于对相同词形所表达的概念及概念之间的关系认识不同，对概念的属性及属性之间的关系认识不同。

也就是说，本文所说的知识在具体表示层面上定义为词形及其关联关系。利用词形的关联关系表达概念、属性、关系。例如：“计算机”这一概念能够被这个词形和词形集合“电脑、软件、硬件、显示器、程序、键盘、计算、上网”等等的关联关系描绘出来，如图 1a

所示。而“盗窃”这一事件能够被这个词形和词形集合“作案、嫌疑人、抓捕、犯罪、警察、受害人、损失”等等的关联关系描绘出来，如图 1b 所示。



a.计算机的词形集合　　b.盗窃的词形集合

图 1 词形集合示例

每一个词形通过与其他词形发生关系而反映这一词形所表达的概念，并且这个关系具有权值来反映关系的不同，从而表达出概念不同。

因此，从整体来看，本文提出的知识的粗略表示即是所有的词形以及带有权值的词形关系图。其数学描述如下：

$$K = （W，R）$$　　（公式 3.1）

其中，W 为所有词形的集合，R 为关系的集合，具体可表示为$r = (w_i，w_j，v)r \in R$，w_i和w_j表示不同的词形，v 表示关联关系的权值。

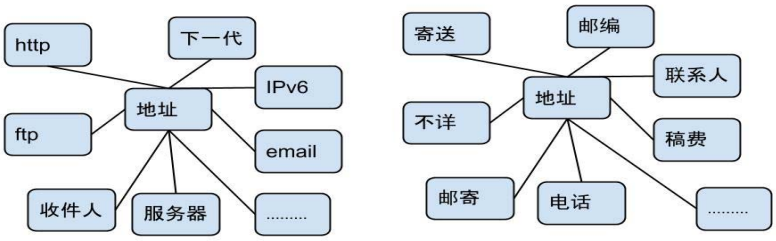
关系图可以转化为关系矩阵，那么每一个词形所代表的概念就可以表示为一个向量：

$$w_i：（v_1，v_2，....v_k，....v_n）$$　　（公式 3.2）

其中v_k是w_i与w_k的关联关系权值，$n = |w|$

这样，每个词形所表达的概念以 n 维空间的一个点来表示，属性蕴含在了向量的值中，而概念之间的关系蕴含于 n 维空间点之间的关系。

至此，知识和词形关系图对应，而概念和词形向量对应。每个人所具有的知识可以和一张词形关系图对应，不同的人，关系图不同，每个词形对应的向量不同，通过图能够计算不同的人的知识差异，如图 2 所示；另外，同一个人的知识也会随着时间而变化（学习、遗忘、时代变迁等因素），即词形关系图随时间变化，如图 3 所示。



a.地址的词形集合(信息行业的人)　　b.地址的词形集合(大众)

图 2 不同行业的人的差异

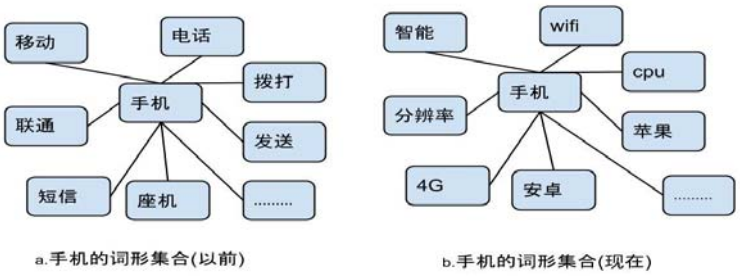


图 3 时间的差异

4 词形关系图的获取与演化

4．1 获取思路

词形关系图的获取依据于人类获取知识的两个特征:

其一是人的知识来源于直接经验和间接经验,而间接经验占据了大多数,在间接经验中,大部分来源于书本。因此根据某个人读过得全部文字内容，应该能够刻画出接近于他的全部知识。

其二是人的知识状态从出生时期的零状态逐渐发展变化,到成年时演变到一种较为稳定的结构状态。

直观来看，词形关系应该源于词的同现关系。同现可以分为词的相邻同现、句子同现、段落同现和文章同现，选取的原则是在一个词的意义影响范围内的同现。相邻同现更多的反映出词的语法关系，而段落同现和文章同现往往会超出该词的意义影响范围，带来更多的噪音，句子作为具有完整意义的语言单位，可以近似的作为每个词的意义影响范围。

对于每一个词形来说，其所对应的词形向量在学习过程中应该类似人的学习过程：即对于知识的不断修正的过程。假设词 w 初始的词形向量为：(v01，v02，....v0k,....v0n)，那么关于 w 知识状态的随时间的变化过程可以表示为以下序列：

$$t_0\sim(v_{01},\ v_{02},\ \ldots v_{0k},\ldots.v_{0n})$$

$$t_1\sim(v_{11},\ v_{12},\ \ldots v_{1k},\ldots.v_{1n})$$

$$t_2\sim(v_{21},\ v_{22},\ \ldots v_{2k},\ldots.v_{2n})$$

$$\vdots$$

$$t_p\sim(v_{p1},\ v_{p2},\ \ldots v_{pk},\ldots.v_{pn})$$

其中 t 表示时刻，t_0是初始时刻，t_p为终止时刻。

这样知识的演变过程就可以量化为每一个词形向量随时间的变化过程,如果这个过程足够长，并且系统读入的文本数量足够多，那么到 t_p 时刻，词形向量应该接近于一个稳定的数值，即 t_p时刻之后再进行学习的话，向量的变化很小。

4．2 获取算法

根据上述分析，本文选定爬山算法实现知识的获取。首先为了计算合理性，每一个词形向量需要归一为单位向量，其演变相当于在一个 n 维空间的球面上寻求一个稳定的点。假设在 t_i时刻词形 w 的词形向量为$\overline{v_i}$，则在 t_{i+1}时刻根据如下方式更新:

$$\overline{v_{i+1}}' = \overline{v_i} + \lambda_k\left(\overline{v_k} - \overline{v_i}\right)\text{（更新词形向量）} \qquad \text{（公式 4.1）}$$

$$\overline{v_{i+1}} = \frac{\overline{v_{i+1}}}{|\overline{v_{i+1}}|}$$ （转变为单位向量） （公式 4.2）

其中 $\overline{v_k}$ 为在 t_{i+1} 时刻新学习到的词形变量， $\lambda_k \in (0,1)$ 为 $\overline{v_k}$ 对 $\overline{v_i}$ 的影响系数，直观如下图所示：

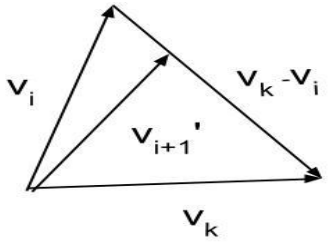


图 4 词形向量更新示意图

在爬山搜索中， λ_k 的值对于算法的效率和结果有很大的影响，其值来源于经验数据，为了公平性原则，在经验值的基础上采用以下公式对其动态修正：

$$\lambda_k = \lambda + 0.1 \times \frac{|w_k| - |w_i|}{|w_i|}$$ （公式 4.3）

其中 λ 为基础更新系数，第二项为修正值，$|w_k|$ 是当前待更新词在语料中出现次数，$|w_i|$ 是当前待更新词在之前的语料中平均出现次数。

4．3 词的选取

选取哪些词作为知识的表示关系到系统的效率和应用效果。本文在初步试验选取了动词和名词，然而有大量的词如“是、有、成为、能够、发生”等等，这些词的功能处于语法和语义之间，同大量的词有很高的共现频率，对于知识表示的贡献甚微，但是不像结构词那样，难以采用列表的方式将其排除。

本文借鉴了人的高层思维的遗忘原理：如果一个词形同太多的词形经常发生关联，那么尽管其同现频率较高，但是在人们阅读或是表达的时候也往往不会注意到。

由此，本文利用艾宾浩斯遗忘曲线[11]对这些词在其他词的词形向量中的权值进行了弱化处理，具体如下：

设 $\overline{v_{mi}}$ 是词 Wm 在 ti 词时刻更新后的词形向量，分别对它的每一个分量进行遗忘，

$$v_{mik} = v_{mik} / e^{\delta \times \alpha_k}$$ （公式 4.4）

其中 k 表示它的第 k 个分量， δ 是基础遗忘系数，来自于经验取值，而

$$\alpha_k = \frac{\overline{v_{ki}} \cdot \overline{v_c}}{|\overline{v_{ki}}| \times |\overline{v_c}|}$$ （公式 4.5）

其中， $\overline{v_{ki}}$ 是 $\overline{v_{mi}}$ 第 K 个分量所对应的词形的词形向量， $\overline{v_c}$ 是一个同每一维夹角相同的词形向量，这样 α_k 就是 $\overline{v_{ki}}$ 和 $\overline{v_c}$ 的夹角余弦，其值越接近于 1，则表示该词关联的词越多且关联值越平均，这样越应该遗忘。

5 实验及结果分析

由于语料所限，本文分别采用了 2012 年 1 月份的人民日报、2012 年 8 月份的羊城晚报作为词形关系来源语料库，分别获得三个学习结果，以期从不同时间段、不同角度学到的知识进行对比。

5．1 收敛性

由于每个人的知识随着年龄的增长逐渐趋于一种较为稳定的状态，那么个性化知识的获取过程也应该随着时间逐步进入一种相对的稳态。因此这里首先关注这种知识表示方案是否会趋于稳定也即算法的收敛性。在这里收敛具体是指随着对词形向量更新次数的增加，词形向量改变的角度趋向于零。具体定义如下：

$$\delta_i = 1 - cos\theta_i = 1 - \frac{\overline{v_i} \cdot \overline{v_{i-1}}}{|\overline{v_i}| \times |\overline{v_{i-1}}|} = 1 - \overline{v_i} \cdot \overline{v_{i-1}} \qquad \text{（公式 5.1）}$$

式中 θ_i 表示第 i 次更新后词形向量同更新前词形向量的角度差，$\overline{v_i}$ 更新后的词形向量，$\overline{v_{i-1}}$ 表示更新前的词形向量，由于它们都是单位向量因此可以化简为右式，我们用 δ_i 表示词形向量的变化量，理论上，$\lim_{i \to \infty} \delta_i = 0$。

图 5 显示了在 2012 年 8 月份的羊城晚报的学习过程中所有词形向量变化量的平均值随更新次数的变化曲线图。其中横轴为知识更新次数（这里设每天更新一次），纵轴为所有词形向量变化量的平均值。从图中可以看出，递减开始较为明显，后续总体呈下降趋势，但斜率变小并且有反复，和人的认知过程较为一致。

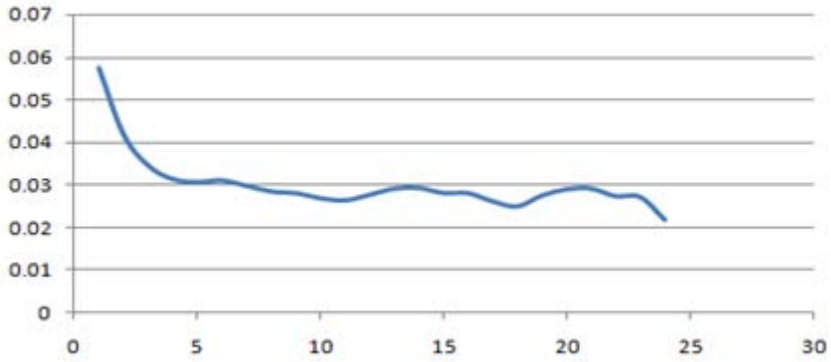


图 5 词形向量变化量的平均值随更新次数的变化

具体到不同的词形，通过观察发现，词形向量变化量随更新次数有三种不同的形态，我们从每一类中分别挑选了一个词形，具体变化如图 6 所示。

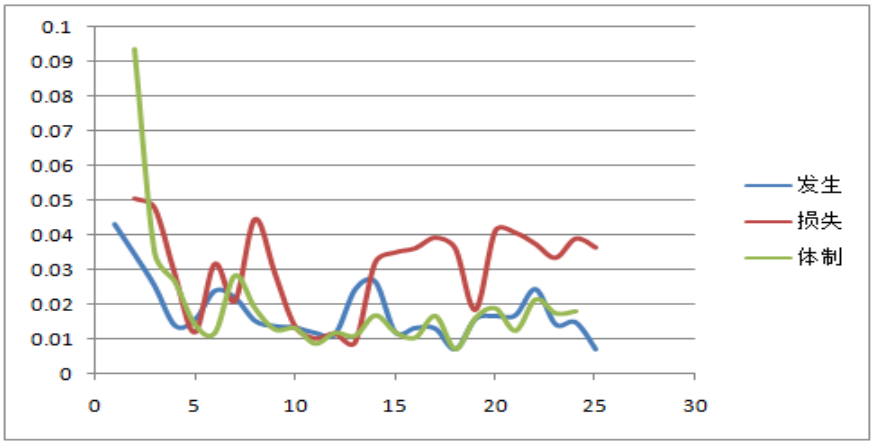


图 6 “发生”、”损失”和”体制”的收敛形态

从图中可以看出，“体制”收敛较快且后续波动较小，而“发生”虽然也能较快收敛但是后续波动相对较大，“损失”一直呈现波动较大的状况。

我们认为“体制”属于领域特征明显的词，其关联的词局限性较大，因此能够较快收敛；“损失”涉及领域较多，如盗窃、灾害、贸易等等，因此短时间内难以快速收敛；“发生”属于使用频率极高的实词，几乎会出现在所有领域，其词形向量方向对每一维趋于平均，因此开始能够快速收敛，但随着新的词的出现会发生一定波动。如果能够有效的鉴别这三种类型的词形向量，对于提高学习的效率和效果都会有所帮助。

5．2 知识差异性

由于一个月的语料所限，大多数词形向量还未达到或接近收敛状态，难以采用统计数据表示知识的差异，我们在此仅仅对比 2012 年 1 月份的人民日报和 2012 年 8 月份的羊城晚报学习到有意义的部分结果，如下表所示：

表 5-1：不同语料的词形向量差异性

词形	相似度	关联词
流入率	1	流出 额 率 指数 流入 板块
		流出 额 流入 指数 率 板块
红十字	0.3	救护 干细胞 捐献 造血 人道 献血 明确 遗体 眼库
		公信力 商 商会 人道 救护 款物 招标 下级 博爱
持卡人	0.1	隐蔽性 刷 窃取 密码 第三人 信用卡 盗 ATM 机 负有 绑定
		VISA AVIS 信用卡 名品 优选 专属 商户 折扣 刷卡 在线
猝死	0.1	体育迷 每逢 盛事 熬夜 心源性 小心 暴发性 心肌炎 世界杯 心肌梗死
		心肌梗死 漏诊 误诊 痛性 知觉 胸口 心梗 冠心病 糖尿病 心血管
撞击	0	脑部 晕厥 生还 箱梁 东西 伸直 齐唱 熔岩 潜规则 水池
		月球 抛射 重力场 内核 外壳 盆地 地貌 天体 火星 演化
地震	0	助理员 医疗队员 护理部 元帅 担架员 伤病员 水灾 震源 震中 网球馆
		有感 震度 捐物 天灾 救援队 防震 余震 地震群 喷发 火山
地址	0	食品店 同城化 邮编 相片 编辑部 总机 发行部 广告部 观后感
		IPv6 代名词 版本 下一代 http 邮箱 收件人 键 快件 卡号

注：表中关联词按照权值从大到小排列的前十个选取，第一行来源于羊城晚报，第二行来源于人民日报，相似度通过集合的 Dice 系数计算得来。

从表中可以看出，除了特定领域的词如“流入率”外，大多数学习结果差异较大，表示了不同角度的认知，如“红十字”，在羊城晚报中和各类捐献相关，而人民日报多指红十字会的运作，同样的情形发生于较低相似度的词形，如果我们假设人民日报的结果代表一个仅仅每天阅读人民日报的人的知识，而羊城晚报的结果代表一个仅仅每天阅读羊城晚报的人的知识，那么可以看出个性化知识的明显差异，直接用于利用词形扩展技术实现个性化搜索结果必将不同。

总之，从初步实验结果可以看出，本文所采用的词形关系图获取方法可行，并且结果也能够粗略反映出用户所拥有的知识，并且表达出不同用户的不同之处。

如何将这种知识表示方法融入当前的应用，如个性化搜索或推荐技术，还有待考证。虽然采用简单的向量扩展方法就可以利用词形关系图进行个性化搜索，但是要以实验同现有技术对比，还存在如下问题：其一是当前的个性化搜索基于用户的搜索历史，个性化推荐基于用户的行为历史如购买、关注等，而本文要基于用户所有的阅读历史，这种较为准确的语料目前还难以获取，假设以不同领域语料作为不同的人的阅读历史，那么结果显而易见，导致无法说明问题。其二，个性化搜索或推荐并无标准的评测平台，实验结果评测的主观性较大，因此这种情况下做实验对比意义不大。

6 总结和展望

本文提出了一种知识表示方法即词形关系图，作为每个人所拥有的知识的粗略表示，目标在于个性化的搜索、话题检测与跟踪、兴趣推荐等应用。同时给出了基于爬山算法的知识获取方法，并取得了初步的结果。从结果可以看出，这种知识表示方法不但可以用于基于海量信息的个性化应用，还可以用于分析媒体之间的差异，或者分析词形的时空变化，考察个人之间知识结构差异等。如果这种获取方法能够将所有的信息作为知识来源的话，那么将得到一个粗略的常识表示。

对于应用来讲，目前获得的结果需要一定的精简，并且需要转换为更好地表示形式如社团结构的图，才能在应用中高效的利用；同时由于这种获取方法数据量巨大，目前的机器配置运行缓慢，需要对获取算法加以优化和改进，最主要的是如何鉴别大量的无用的更新信息，将是实用性的关键一步；另外，在实验中我们发现词形同领域的关系导致其收敛性的差别，如何有效的分析和利用这种差别也是我们下一步将要研究的内容；当前我们还无法评估这种表示方法在具体应用上的效果，因此如何将这种知识表示方法融入现有的个性化应用技术，并能够进行有效的评估，也是进一步要研究的重点内容。

参考文献：

[1]秦兵,刘挺,李生. 多文档自动文摘综述[J]. 中文信息学报. 2005, 19(6): 13-20.

[2]Agrawal R, Gollapudi S, Halverson A, et al. Diversifying search results[C]. Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York, 2009: 5-14.

[3]朱岩，林泽楠. 电子商务中的个性化推荐方法评述[J]. 中国软科学. 2009,2: 183-192.

[4] Guy I., Zwerdling N., Ronen I., et al. Social media recommendation based on people and tags[C]. Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. Switzerland, 2010 : 194-201.

[5] George A. Miller. the WordNet project[DB].[2012-12-27]. http://wordnet.princeton.edu/

[6] 董振东, 董强. 知网[DB]. [2013]. http://www.keenage.com/.

[7] Ruppenhofer J, Ellsworth M, Petruck M R L, et al. FrameNet II: Extended Theory and Practice. http://framenet.icsi.berkeley.edu/.

[8] Weischedel R, Pradhan S, Ramshaw L, et al. OntoNotes Release 4.0[DB].[2013]. http://www.bbn.com/NLP/OntoNotes/.

[9] 周强, 王俊俊, 陈丽欧. 构建大规模的汉语事件知识库[J]. 中文信息学报, 2012, 26(3): 86-91.

[10] Passant, A. Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs[C]. International Conference on Weblogs and Social Media, Boulder, Colorado, March 2007.

[11] HERMANN E. Memory: A contribution to experimental psychology [M]. New York: Columbia Teachers College, 1913.