基于维基百科社区挖掘的词语语义相似度计算

彭丽针 吴扬扬
(华侨大学计算机科学与技术学院 厦门 361021)

摘 要 词语语义相似度计算在自然语言处理如词义消歧、语义信息检索、文本自动分类中有着广泛的应用。不同于传统的方法,提出的是一种基于维基百科社区挖掘的词语语义相似度计算方法。本方法不考虑单词页面文本内容,而是利用维基百科庞大的带有类别标签的单词页面网信息,将基于主题的社区发现算法 HITS 应用到该页面网,获取单词页面的社区。在获取社区的基础上,从3个方面来考虑两个单词间的语义相似度:(1)单词页面语义关系;(2)单词页面社区语义关系;(3)单词页面社区所属类别的语义关系。最后,在标准数据集 WordSimilarity-353 上的实验结果显示,该算法具有可行性且略优于目前的一些经典算法;在最好的情况下,其 Spearman 相关系数达到 0.58。

关键词 语义相似度,社区发现,维基百科
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.009

Semantic Similarity Computing Based on Community Mining of Wikipedia

PENG Li-zhen WU Yang-yang
(School of Computer Science & Technology, Huaqiao University, Xiamen 361021, China)

Abstract Words semantic similarity computing has been widely used in natural language processing, such as word sense disambiguation, information retrieval, text auto categorization. Different from traditional methods, we presented an algorithm based on community mining of Wikipedia to compute words semantic similarity. Our method makes use of the huge Wikipedia page network with category labels rather than its textual content. To get the community of a word page, we applied the HITS, which is a community discovery algorithm based on the theme, to pages network. Based on the gotten community, we measured the semantic similarity between two words from three aspects: (1) semantic relations between the two word pages, (2) semantic relations between the two communities of word page, (3) semantic relations between the categories which two communities belong to. Finally, tests on standard data sets WordSimilarity-353 show that the method we proposed is feasible and slightly better than some classic algorithms. In the best case, the Spearman correlation coefficient reaches 0.58.

Keywords Semantic similarity, Community discovery, Wikipedia

词语语义相似度计算在自然语言处理中有着很广范的应用,例如文本摘要自动提取、词义消歧、语义信息检索、文本自动分类、机器翻译、查询扩展等。

现有的词语语义相似度计算方法主要有两大类。

(1)基于大规模语料库进行统计,依据词汇上下文信息的分布进行计算的方法。这类方法能够对词汇间的语义相似性进行比较精确和有效的度量,但需要依赖于训练所用的语料库,计算量大,计算方法复杂。

(2)基于语义词典的计算方法,常见的词典有 WordNet、HowNet 及《同义词词林》,如刘群等人[1]提出的基于《知网》的词语相似度计算,Leacock 等人[2]提出的基于 WordNet 的路径方法(简称 lch),Resnik[3]提出的基于 WordNet 的信息论方法(简称 res)等。虽然这类方法相对更简单有效,但是这些语义词典库的更新很难跟上如今互联网上新概念、新词语不断涌现的速度,无法做到收录现实应用中所有的词语。而近年来随着维基百科的兴起,人们发现这种由互联网用户自由贡献、共同协作的方式构建的大规模知识资源是一种很实用的语料库。

目前基于维基百科的词语相似度计算方面的研究也比较多。比较经典的有 Michael Strube 等人[4]提出的 WikiRelate! 方法,其借鉴 WordNet 的词语语义相关度计算方法,将 WordNet 的概念层次结构改用为维基百科的文档类型结构,利用维基百科的文档内容代替 WordNet 的词汇定义,进行词汇语义相关性计算,效果与 WordNet 度量方法相当;Evgeniy Gabrilovich 等人[5]提出的 ESA 算法根据输入的词语获得维基百科的主题页面,将这个词语的语义映射到由维基百科概念所形成的带有权重的高维向量空间中,再利用向量空间模型计算两个向量之间的余弦相似度来得到两词汇的语义相关度,但该方法所要处理的文本繁重,计算开销大。David Milne[6]纯粹利用 Wikipedia 的链接结构进行词汇语义相关性

到稿日期:2015-06-10 返修日期:2015-07-21 本文受福建省科技计划重点项目(2011H0028)资助。
彭丽针(1990—),女,硕士生,主要研究方向为自然语言处理、数据挖掘,E-mail:18750917326@163.com;吴扬扬(1957—),女,教授,硕士生导师,主要研究方向为数据管理、数据挖掘。

的度量，其效果介于 WikiRelate! 和 ESA 之间；Rui-Qin[7] 等人为了避免繁重的文本处理工作，只考虑链接结构，并赋予链接不同的权重；Feiyue Ye 等人[8] 同时考虑页面网络和类别网络来计算语义相似度；Taieb 等人[9] 结合两个单词在维基百科中不同模块（如文章、类别、维基百科类别图等）的语义信息来计算其语义相关性；盛志超等人[10] 利用页面的链接信息，通过模仿人类联想的方式，考虑词语的类别信息来计算两个词语的相似度；孙琛琛等人[11] 提出一种基于维基百科的文章网络和分类树的结构信息来计算词语的语义关联度；刘晓亮[12] 将典型的随机游走模型 PPR（Personalized PageRank）应用到维基语义图，计算词语间的语义相关度。

维基百科中存在社区现象[13,14]，每个社区由主题相关的维基百科单词页面所组成，并倾向于表示同一主题。因此，本文提出了一种基于维基百科社区挖掘的词语语义相似度计算方法。为了减少繁杂的文本处理，本文不考虑单词页面文本内容，仅利用维基百科丰富的页面链接结构和类别标签来计算词语的语义相似度。首先，构造带有类别标签的单词页面网；然后，将基于主题的社区发现算法 HITS 应用到该页面网，获取单词的维基百科页面社区。在获取社区的基础上，从 3 个方面来考虑两个单词间的语义相似度：1）两个单词页面本身是否存在链接关系，根据维基百科页面的特性[15]，若存在链接，则一定程度上说明这两个单词存在语义关系。2）两个单词的维基百科页面社区成员之间的语义关系。由于社区里的页面不仅反映了一个共同主题，同时也是对该单词的一种间接解释，因此若其社区之间互相链接频繁，则一定程度上说明这两个单词存在语义关系。3）单词页面社区里的成员隶属的类别的语义关系。由于用户在编辑单词页面时会根据其主题赋予单词页面一些语义标签即类别，且一个单词页面可以属于不同的类别，一个类别下面可以有多个单词页面，因此本文分别获取两个单词页面社区成员所属的类别，将获取的类别表示成向量空间模型，通过计算向量之间的余弦相似度来判断其语义关系。

1 构建维基百科带有类别标签的单词页面网

后文将由维基百科单词解释页面构成的网络简称为页面网，可用有向图 $G\langle V,E\rangle$ 来表示，其中，V 为节点集合，V 的每一个节点代表一个带有类别标签的单词页面，如式（1）所示：

$$V=\{(word_1,c_1,c_2,\cdots,c_i),(word_2,c_1,c_2,\cdots,c_j),\cdots,(word_n,c_1,c_2,\cdots,c_k)\} \tag{1}$$

其中，n 为节点的个数，$c_1,c_2,\cdots,c_i$ 是单词页面的类别标签，$i,j,\cdots,k$ 为单词页面的类别标签的数量。

E 为有向边集合，E 中的边代表从一个单词页面到另一个单词页面的链接关系。

根据维基百科页面的特性，单词页面之间存在丰富的反映语义关系的引用链接。考虑到每个单词页面内容的布局，包括单词解释、语源学、历史等，若是对所有模块的链接都加以考虑，可能会出现语义偏移问题。对此，这里仅提取页面结构中解释模块的单词链接。

构建页面网 G 的基本步骤包括：

（1）从原始的维基百科数据集中提取单词解释页面的标题（即单词）、页面解释模块的链出/链入链接和页面所属的类别标签，构建节点集合 V。

（2）执行算法 1，构建维基百科单词解释页面网络 G。

算法 1 构建维基百科单词解释页面网络 G

输入：单词解释页面节点集合 V

输出：单词解释页面网络 G〈V,E〉

方法：

(1)将节点集合 V 存储在一维数组 Vertex 中；将节点个数|V|赋给 N；

(2)初始化边 Edge[N][N]；\\有向边的集合用邻接矩阵表示

(3)For(int i=0;i<Vertex. length();i++)

```
    For(int j=0;j<Vertex. length();j++){
      If(节点 i 到节点 j 存在链接)
        Edge[i][j]=1;
      Else
        Edge[i][j]=0;
    }
```

(4)返回页面网络 G〈V,E〉。

2 单词的维基百科社区发现

2009 年，Lizorkin 等人[14] 对维基百科中的社区进行分析，并证实其存在社区现象，且每个社区由主题相关的维基百科文章所组成，并倾向于表示同一主题。

本文应用经典的基于主题的社区发现算法——HITS[16]，在上述构建的页面网中发现指定单词所在的社区。具体步骤如下：

（1）输入一个单词。

（2）获取根集：在页面网中获取该单词的链出和链入页面。

（3）获取基集：对根集进行一层扩展，包括根集的链出和链入页面。

（4）迭代计算页面的 Authority 权威值和 Hub 中心值：

A）给每个单词页面赋予两个度量值：中心值 h_i 和权威值 a_i。初始化向量，$a=(a_1,a_2,\cdots,a_n)$ 和 $h=(h_1,h_2,\cdots,h_n)$，初始值均为 1。

B）采用式（2）和式（3）分别计算页面的 Authority 值和 Hub 值。

$$A_p=\sum_{q->p}H_q \tag{2}$$

$$H_p=\sum_{p->q}A_q \tag{3}$$

C）分别对页面的权威值和中心值从大到小排序，分别选取排名前 10 的页面作为维基百科社区。

3 基于维基百科单词页面社区的语义相似度计算方法

前面一节用经典的 HITS 算法获取某个单词页面所在的社区。基于单词页面社区，下面从 3 个方面来考虑两个单词间的语义相似度：1）单词页面语义关系；2）单词页面社区语义关系；3）单词页面社区所属类别的语义关系。

3.1 单词页面语义关系 *Sim*1

根据两个单词页面解释模块的引用链接情况，分 3 种情况处理：1）若两个单词存在互相链接，说明其语义相关性很强，赋予其值为 1；2）若两个单词只有单一的链接，其语义关系没有前者那么强，赋予其值为 0.5；3）若两个单词不存在链接，其语义关系较弱，赋予其值为 0。因此，两个单词页面的 *Sim*1 值由式（4）计算。

$$Sim1(Word_1,Word_2)=\begin{cases}1, & \langle V_{word_1},V_{word_2}\rangle\in E \text{ and } \langle V_{word_2},V_{word_1}\rangle\in E\\ 0.5, & \langle V_{word_1},V_{word_2}\rangle\in E \text{ or } \langle V_{word_2},V_{word_1}\rangle\in E\\ 0, & \langle V_{word_1},V_{word_2}\rangle\notin E \text{ and } \langle V_{word_2},V_{word_1}\rangle\notin E\end{cases} \quad (4)$$

3.2 单词页面社区语义关系 *Sim2*

根据前面对单词页面社区的分析，社区里的单词倾向于一个共同主题。若两个社区之间存在的链接比较频繁，则一定程度反映了其存在语义关系。社区语义关系 *Sim2* 由式(5)计算。

$$Sim2(Word_1,Word_2)=\frac{N}{2*N_1*N_2} \quad (5)$$

其中，N 为社区间存在的实际边数；N_1、N_2 分别为单词 1 和单词 2 所在社区所包含的页面数；$2*N_1*N_2$ 为社区间可能存在的最多边数。

社区间链接的边数越多，社区语义关系 *Sim2* 的值就越大；如果社区间链接的边数为 0，其语义关系 *Sim2* 的值为 0。

3.3 单词页面社区成员隶属的类别的语义关系 *Sim3*

类别是单词页面的语义标签。因此，考虑将两个单词所在社区成员隶属的类别集合用向量空间模型表示，然后通过余弦相似度来判断其语义关系。

(1)单词所在社区成员隶属的类别集合

假设 $Word_1$、$Word_2$ 的社区成员隶属的类别集分别为式(6)、式(7)：

$$S(Word_1)=\{c_{11},c_{12},\cdots,c_{1n}\} \quad (6)$$

$$S(Word_2)=\{c_{21},c_{22},\cdots,c_{2m}\} \quad (7)$$

其中，$c_{11},c_{12},\cdots,c_{1n}$ 和 $c_{21},c_{22},\cdots,c_{2m}$ 分别为单词 1 和单词 2 社区成员所属的类别。类别的并集表示为：$S(Word)=\{c_1,c_2,\cdots,c_i\}$，$i$ 是类别并集的项数。

(2)类别集合的向量表示

由于前面两个类别集的大小不同，其用向量空间模型表示的类别向量的维度也会不同。为了便于向量的余弦相似度的计算，本文考虑将两个类别向量映射到其并集向量。

类别向量表示为 $V(Word_k)=(w_{k1},w_{k2},\cdots,w_{ki})$。其中，$k$ 表示单词 k，取值为 1 或 2；i 是类别并集 $S(Word)$ 中类别的数量；w_{ki} 是 $S(Word)$ 中类别 c_i 的权重。

(3)权重 w_{ki} 的计算

w_{ki} 由式(8)计算，取类别 c_i 与单词 k 的社区成员隶属的类别的最大相似度值。

考虑到维基百科中类别的表示常常是两个单词的组合型，如 Economic_anthropology，此处用编辑距离公式[19]来计算相似度。最后，对整个向量的权重做归一化处理。

$$W_{ki}=\mathrm{Max}\{LD(c_i,c_{k1}),LD(c_i,c_{k2}),\cdots,LD(c_i,c_{kn})\} \quad (8)$$

其中，$c_{k1},c_{k2},\cdots,c_{kn}$ 为单词 k 的社区成员隶属的类别；c_i 为类别；$LD(c_i,c_{kn})$ 代表编辑距离计算函数。

(4)余弦相似度

$$Sim3(Word_1,Word_2)=\cos\theta=\frac{\sum_{k=1}^{i}(w_{1k}(V(Word_1))*w_{2k}(V(Word_2)))}{\sqrt{(\sum_{k=1}^{i}w_{1k}^2(V(Word_1)))}*\sqrt{(\sum_{k=1}^{i}w_{2k}^2(V(Word_2)))}} \quad (9)$$

其中，i 为类别并集的项数；$w_{1k}(V(Word_1))$ 和 $w_{2k}(V(Word_2))$ 分别为 $Word_1$、$Word_2$ 的社区成员隶属的类别向量中第 k 个类别 c_k 的权重。

综上，从 3 个方面来计算两个单词的语义相似度，基于维基百科社区挖掘的词语语义相似度由式(10)计算。

$$Sim(Word_1,Word_2)=\alpha*Sim1+\beta*Sim2+\gamma*Sim3 \quad (10)$$

其中，α、β、γ 为 *Sim*1、*Sim*2、*Sim*3 的权重。

4 实验及其结果分析

4.1 构建维基百科带有类别标签的页面网

维基百科是一个免费的、开放的、自由的在线百科全书。其中的每一个词语都由一个对应的页面文档来解释该单词的含义且页面的标题为该单词。在解释页面中，存在很多链接，链接对应的是另一个单词的解释页面。解释页面和解释页面之间的链接可以反映对应的主题词语之间的相关性。

维基百科中存在很多种页面，如单词解释页面、类别页面、消歧义页面等，其中解释页面之间存在链接关系，类别页面之间显示层次关系，而解释页面与类别页面之间也存在所属关系。

本文采用的英文维基百科数据集版本是 enwiki-20141106[20]，并针对性地选用其中的 3 个文件，即

(1)enwiki-20141106-pages-articles. xml. bz2；

(2)enwiki-20141106-pagelinks. sql. gz；

(3)enwiki-20141106-categorylinks. sql. gz。

其中，文件(1)是单词解释页面内容的 xml 文件，包括 title、body 等标签内容；文件(2)是单词解释页面的链接关系文件，包括单词解释页面间的引用链接及单词解释页面到类别页面的链接；文件(3)是类别页面的链接关系文件，包括类别页面基本信息、页面间的层次关系及类别页面下的单词解释页面。

调用 Wikipedia 的 DataMachine[21]工具包将原数据集结构化成各个表的数据文本文件，如 category. txt、page. txt、page_outlinks. txt 等，其中页面数量达到 217 多万条，类别数量达到 117 多万条，页面间链接数近 31500 万个。

在 Mysql 数据库中构建对应表结构，并将数据存储到数据库中。执行算法 1 构建带有类别标签的单词解释页面网络。

4.2 单词的维基百科社区发现结果分析

根据前文提到的单词的维基百科社区发现方法，假设输入单词为 Money，通过社区挖掘，所得的社区如图 1 所示，其中，每个节点为一个单词页面，经过分析，该社区内部页面之间链接较频繁，且社区内倾向于一个共同主题 economics。

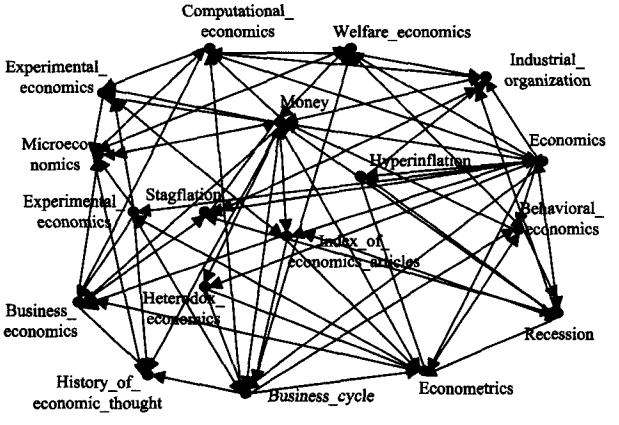

图 1 单词 Money 的维基百科社区

4.3 词语语义相似度计算结果分析

本文采用经典的英文词对集 WordSimilarity-353 测试集进行实验，该数据集包括 353 个词对，且每个词对均有一个人工判断值，该值是由 13～16 个人进行判断后取平均值所得。表 1 记录了不同算法在 WordSimilarity-353 数据集上的部分运行结果(前 15 个词语对)。其中，Human 是原数据集人工获得的相似度；Lch 是 WikiRelate[4] 中的 lch 算法获得的结果；CWPNRelate 是文献[10]中的方法的实验结果；Rel 是文献[8]的方法的实验结果；WPCRelate 是本文提出的基于维基百科社区挖掘的词语语义相似度算法获得的结果，根据实验过程，取 $\alpha=0.3$、$\beta=0.2$、$\gamma=0.5$。此外，为了便于与人工判断值进行比较，本文统一将所有数值进行归一化处理，使得其值均落在[0,10]内。

表 1 WordSimilarity-353 前 15 个词对的结果

word pair	Human	Lch	CWPN Relate	Rel	WPC Relate
love-sex	6.77	7.35	6.58	4.92	5.78
tiger-cat	7.25	4.14	5.99	6.26	7.49
book-paper	7.46	9.13	9.30	4.34	5.07
computer-keyboard	7.62	8.57	5.83	7.24	6.28
computer-internet	7.58	6.38	7.51	4.38	6.30
plane-car	5.77	7.02	5.27	4.80	5.30
train-car	6.31	9.22	4.90	2.67	7.19
telephone-communication	7.50	7.01	5.83	2.76	6.88
television-radio	6.77	7.67	5.21	4.45	6.46
media-radio	7.42	8.43	5.77	3.40	5.40
drug-abuse	6.85	6.71	5.30	3.30	5.99
bread-butter	6.19	7.40	4.83	4.63	6.39
cucumber-potato	5.92	4.73	5.62	8.41	7.39
doctor-nurse	7.00	8.57	5.92	3.31	7.27
professor-doctor	6.62	8.57	7.54	3.55	5.95

此外，为了更好地将所得结果与不同的算法进行对比，本文还采用 Spearman 相关系数将计算结果与人工结果进行比较。其中，若 Spearman 相关系数越大，则说明该算法的计算结果越接近人工判断的结果。结果如表 2 所列，其中，WordNet 是一组基于 WordNet 的经典词语相似度算法所得的结果范围[3]，如 lch、res、lesk 等；WikiRelate! 是将前面基于 WordNet 的算法转移到 Wikipedia 中计算获得的结果范围[4]；Roget's Thesaurus 表示基于 Roget's Thesaurus 所得的结果[17]；LSA 表示 LSA 算法所取得的结果[18]；CWPNRelate 和 Rel 分别是文献[10]和文献[8]的实验结果；WPCRelate 表示本文算法所得结果。

表 2 Spearman 相关系数的比较结果

Algorithm	Correlation with humans
WordNet	0.21～0.35
WikiRelate!	0.19～0.48
Roget's Thesaurus	0.55
LSA	0.56
CWPNRelate	0.5342
Rel	0.53
WPCRelate	0.58

经过与一些经典的算法以及近几年同是考虑维基百科页面网和类别网的不同方法比较，可以看出本文提出的方法在该数据集上的实验结果略优于这些算法，更接近人工判断结果，说明了其具有可行性和有效性。

结束语 本文提出一种基于维基百科社区挖掘的词语语义相似度计算方法。实验结果表明，该方法计算的结果略优于几个经典算法以及近年来同类语义相似度算法，更接近人工判断结果，说明了本文的方法具有一定的可行性和有效性。

下一步的研究方向：1)本文在计算单词页面语义关系时，从计算的简便性出发，单纯考虑其互相是否链接的这种情况，下一步考虑适当抽取其页面部分内容，综合考虑其链接及内容的语义关系；2)考虑到维基百科中除了丰富的单词页面链接结构外，还存在丰富的类别层次关系，在计算单词页面社区所属类别间的语义关系时，将类别间的层次关系考虑进来。3)本文对计算结果尚未做深入比较分析，未考虑到该方法的适用范围，下一步将对这方面做深入研究。

参 考 文 献

[1] Liu Qun, Li Su-jian. Word Similarity Computing Based on Hownet[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76(in Chinese)
刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学，2002，7(2)：59-76

[2] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification[M]// WordNet: An Electronic Lexical Database. 1998: 265-283

[3] Resnik P. Using information content to evaluate semantic similarity in a taxonomy[J]. arXiv: cmp-lg/9511007, 1995

[4] Strube M, Ponzetto S P. WikiRelate ! Computing semantic relatedness using Wikipedia[C]// AAAI. 2006: 1419-1424

[5] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis[C]// IJCAI. 2007: 1606-1611

[6] Milne D. Computing semantic relatedness using wikipedia link structure[C]// Proceedings of the New Zealand Computer Science Research Student Conference. 2007: 63-70

[7] Wang Rui-qin. Measurement of Semantic Relatedness between Words Based on Link Information of Wikipedia[J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(4): 385-389(in Chinese)
王瑞琴. 基于 Wikipedia 链接信息的词汇语义相关性度量[J]. 情报学报，2013，32(4)：385-389

[8] Ye F, Zhang F, Luo X, et al. Research on measuring semantic correlation based on the Wikipedia hyperlink network[C]// 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). IEEE, 2013: 309-314

[9] Taieb M A H, Aouicha M B, Hamadou A B. Computing semantic relatedness using Wikipedia features[J]. Knowledge-Based Systems, 2013, 50: 260-278

[10] Sheng Zhi-chao, Tao Xiao-peng. Semantic Similarity Computing Method Based on Wikipedia [J]. Computer Engineering, 2011, 37(7): 193-195(in Chinese)
盛志超，陶晓鹏. 基于维基百科的语义相似度计算方法[J]. 计算机工程，2011，37(7)：193-195

[11] Sun Chen-chen, Shen De-rong, Shan Jing, et al. WSR: A Seman-

tic Relatedness Measure Based on Wikipedia Structure[J]. Chinese Journal of Computers, 2012, 35(11): 2361-2370(in Chinese)
孙琛琛,申德荣,单菁,等. WSR:一种基于维基百科结构信息的语义关联度计算算法[J]. 计算机学报,2012,35(11):2361-2370

[12] Liu Xiao-liang. Research on Computation of Lexical Semantic Relatedness Based on Wikipedia Semantic Graph[J]. Journal of the China Society for Scientific and Technical Information, 2014,33(11):1124-1132(in Chinese)
刘晓亮. 基于维基语义图的词语语义相关度计算研究[J]. 情报学报,2014,33(11):1124-1132

[13] Bellomi F, Bonato R. Network analysis for Wikipedia[C]// Proceedings of Wikimania. 2005

[14] Lizorkin D, Medelyan O, Grineva M. Analysis of community structure in wikipedia[C]// Proceedings of the 18th International Conference on World Wide Web. ACM, 2009:1221-1222

[15] Li Yun. Research about semantic knowledge mining based on the Chinese Wikipedia[D]. Beijing: Beijing University of Posts and Telecommunications, 2009

[16] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604-632

[17] Jarmasz M. Roget's thesaurus as a lexical resource for natural language processing[J]. arXiv: 1204. 0140, 2012

[18] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25(2/3): 259-284

[19] Levenshtein V I. Binary codes capable of correcting deletions, insertions, and reversals[C]// Soviet Physics Doklady, 1966, 10(10): 707-710

[20] 维基百科数据集[OL]. http://dumps. wikimedia. org/

[21] DataMachine[OL]. http://search. maven. org/#search|ga|1|tudarm stadt. ukp

(上接第 15 页)

[38] Shi Y, Karatzoglou A, Baltrunas L, et al. TFMAP: optimizing MAP for top-n context-aware recommendation [C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012: 155-164

[39] Wang Peng, Jing Li-ping. Improved One-class Collaborative Filtering for Recommendtion System[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(10): 1231-1238(in Chinese)
王鹏,景丽萍. 改进的单类协同过滤推荐方法[J]. 计算机科学与探索,2014,8(10):1231-1238

[40] Li Gai, Li Lei. One-class collaborative filtering based on matrix factorization[J]. Application Research of Computers, 2012, 29(5): 1662-1665(in Chinese)
李改,李磊. 基于矩阵分解的单类协同过滤推荐算法[J]. 计算机应用研究,2012,29(5):1662-1665

[41] Núñez-Valdéz E R, Lovelle J M C, Martínez O S, et al. Implicit feedback techniques on recommender systems applied to electronic books[J]. Computers in Human Behavior, 2012, 28(4): 1186-1193

[42] Li Gai, Li Lei. One-class collaborative filtering algorithm based on social network[J]. Journal of Hubei University(Natural Science), 2014, 36(4): 333-338(in Chinese)
李改,李磊. 基于社交网络的单类协同过滤算法[J]. 湖北大学学报(自然科学版),2014,36(4):333-338

[43] Luo Sheng-mei, Lin Yun-zhen, Ye Xiao-long, et al. One Class Collaborative Filtering Algorithm Based on Transfer Learning [J]. Hans Journal of Data Mining, 2013, 3(1): 12-17(in Chinese)
罗圣美,林运祯,叶小伟,等. 基于迁移学习的单类协同过滤算法[J]. 汉斯出版社-数据挖掘,2013,3(1):12-17

[44] Rendle S. Factorization machines with libFM[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(3): 57

[45] Shi Y, Karatzoglou A, Baltrunas L, et al. Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering [C] // Proceedings of the Sixth ACM Conference on Recommender Systems. ACM, 2012: 139-146

[46] Shi Y, Karatzoglou A, Baltrunas L, et al. xCLiMF: optimizing expected reciprocal rank for data with multiple levels of relevance [C]// Proceedings of the 7th ACM Conference on Recommender Systems. ACM, 2013: 431-434

[47] Zhu Yu-xiao, Lv Lin-yuan. Evaluation metrics for recommender systems[J]. Journal of Electronic Science and Technology of China, 2012, 41(2): 163-175

[48] Herschtal A, Raskutti B. Optimising area under the ROC curve using gradient descent [C]// Proceedings of the Twenty-first International Conference on Machine Learning. ACM, 2004: 49

[49] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval [M]. Cambridge: Cambridge University Press, 2008

[50] Craswell N. Mean Reciprocal Rank [M]// Encyclopedia of Database Systems. Springer, 2009: 1703-1703

[51] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering [C]// Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1998: 43-52

[52] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents [C]// Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2000: 41-48

[53] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques [J]. ACM Transactions on Information Systems (TOIS), 2002, 20(4): 422-446