基于维基百科的汉越词语相似度计算

杨启悦[1],余正涛[1],洪旭东[1],高盛祥[1],汤智文[2]

(1. 昆明理工大学 信息与自动化学院,云南 昆明 650500; 2. 北京航空航天大学 计算机科学与工程学院,北京 100191)

摘 要: 为了解决跨语言汉越词语相似度计算问题,以维基百科多语言概念页面作为桥梁,利用概念之间存在的翻译对应关系、词语出现在不同概念页面及与其他概念之间存在共现关系,提出了基于维基百科的汉越词语相似度计算方法,该方法首先提取维基百科中汉语越南语具有对应关系的概念集合,构建双语概念特征空间,然后根据词语在相应概念描述文本中出现的词频特征,以及词语与概念在其他概念文本中的共现特征构建词语的概念向量值,最后通过夹角余弦对两个向量进行词语相似度计算。实验结果表明提出的方法在汉越双语词语相似度计算上表现了好的效果,概念共现关系能够提高词语相似度的准确率。

关键词: 汉语; 越南语; 词语相似度; 维基百科; 概念; 共现关系; 对应关系; 词频

中图分类号: TP391.1 **文章编号:** 1005-9830(2016)04-0461-06

DOI: 10.14177/j.cnki.32-1397n.2016.40.04.014

Chinese-Vietnamese word similarity computation based on Wikipedia

Yang Qiyue[1], Yu Zhengtao[1], Hong Xudong[1], Gao Shengxiang[1], Tang Zhiwen[2]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China;
2. School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: In order to solve the word similarity between Chinese and Vietnamese, setting the multi-language concept description page from Wikipedia as a bridge, using translation correspondence between concepts, words appearing in different concept pages, and the co-occurrence relationship between words and other concepts, the method of calculating the similarity between Chinese-Vietnamese words based on Wikipedia is proposed. The set of Chinese-Vietnamese correspondence concept is extracted from Wikipedia to construct bilingual concept feature space. According to the word frequency features appearing in the corresponding concept text, and the co-occurrence features of words and concepts in other concept texts, we construct the concept vector value of words. The similarity between two vectors is calculated by the angle cosine. The experimental results indicate that the proposed method has good

收稿日期: 2015-12-26 **修回日期:** 2016-01-19

基金项目: 国家自然科学基金(61175068,61472168);云南省自然科学重点项目(2013FA030)

作者简介: 杨启悦(1992-),女,硕士生,主要研究方向:自然语言处理,Email: yanghelen412@qq.com; **通讯作者:** 余正涛(1970-),男,博士,教授,主要研究方向:自然语言处理、信息检索、机器翻译,E-mail: ztyu@hotmail.com。

引文格式: 杨启悦,余正涛,洪旭东,等. 基于维基百科的汉越词语相似度计算[J]. 南京理工大学学报,2016,40(4):461-466.

投稿网址: http://zrxuebao.njust.edu.cn

effect on the similarity computation between Chinese and Vietnamese words, and the concept co-occurrence relationship can improve the accuracy of word similarity.

Key words: Chinese; Vietnamese; word similarity; wikipedia; concept; co-occurrence relationship; corresponding relation; word frequency

越南与我国交流密切,文本及语言理解对加强汉越交流越来越重要,词语相似性分析是文本理解与语言理解的基础,汉越双语词语相似度分析对汉越跨语言信息检索、话题发现、文本分类都有重要支撑作用。目前词语相似度计算方法可分为基于知识库或基于大规模词语共现统计信息的相似度计算两类方法,基于知识库的相似度计算方法,主要思想是利用语言中知识库如 HOWNET、WORDNET,借助于词典所表征词语之间的知识结构关系来计算词语之间的相似度,Mostafa 等人提出了一个基于 wordnet 层次结构特征的模型,定义了词语在 wordnet 中层次结构位置的加权公式,通过计算两个词语的距离来判断两个词语的相似性[1]。田永乐等人为解决词语在语义网自适应学习系统中相似度计算的问题,提出根据词语的义项在同义词词林的位置和编码计算词语的相似度[2]。刘群等人通过研究《知网》中知识描述语言的语法,区分词语多个义原的关系在词语相似度计算中所起的作用,从而提出利用《知网》进行词语相似度计算的算法[3]。詹志建等人提出的方法通过计算百科名片、词条正文、开放分类和相关词条部分的相似度加权得到词语相似度[4],张冰怡等人利用词语的相邻关系、编码特征等信息计算词语与知识库中概念之间的相似度[5],王文等人更是基于微博文本信息,通过判断词语之间情感极性,从而计算词语间相似度[6];基于大规模词语共现统计信息相似度计算方法,其主要思想是使用大规模统计语料,统计获得任何两个词语之间的共现信息,借助于共现关系来计算词语之间的相关性,例如,Dagan 等人利用联想词语的分布概率来体现词语之间的相关性,提出了一个基于词语分布的概率模型,运用现有的信息来评估一组词对最大的相似概率[7]。Gracia 等人提出了一个新的词语语义相似度度量方法,该方法利用搜索引擎返回的搜索频率信息计算词语之间的相似度[8]。赵军等人定义了词语关联分布关系,通过词语所在上下文内容计算不同词语的关联度,用以提高词语相似度计算的结果[9]。以上两类相似度计算方法基本上都是在单语环境下进行,在双语词语相似度计算方面,吴思颖等人提出构建一个中英对照 WordNet,借助同义词关系,将要参与计算的词语在词典上查找出同一语种的同义词集,然后计算同义词集之间的相似度,获得跨语言词语相似度[10]。Vulic 等人通过构建一个多语言主题概率模型,利用主题模型生成源语言词语对应的目标语言相似词语概率,选取最大概率作为跨语言词语间相似度[11]。而对于计算多语种环境下的词语相似度计算方法,目前开展的工作还相当有限其主要原因是对于双语环境,目前还没有跨语言的多语言知识库,单纯的互译知识库还不能完全准确体现目标语言特性,另外基于统计的词语相关性计算需要大规模双语的统计语料,平行语料资源的匮乏限制了双语研究的发展。

维基百科作为全世界最大的多语种、开放式的在线百科全书[12],其对不同语言的概念进行描述,而且不同语言之间的概念具有翻译对应关系,仅中文和越南语描述的概念就有 80 多万个,这些概念通过不同的文本来描述,任何一个词语可能在不同概念描述页面中出现,而且任何一个词语与其他概念之间在描述页面中会存在一定共现关系,如“天文学”一词出现在了“国际天文联合会”的概念中,同时“天文学”一词与概念“国际天文联合会”在“矮行星”等概念上存在共现关系。不同语言描述这些概念在不同语种之间存在互译关系。因此,本文作者认为可以利用这些关系及词语统计特性来计算词语相似关系,提出基于维基百科的汉越跨语言词语相似度计算方法,利用维基百科中汉语越互译概念词作为汉越词语相似度计算的桥梁,构建双语概念特征空间,将词语在不同概念描述文本中出现的词频以及与概念在其他文本中的共现次数作为特征值,从而计算汉越词语之间的跨语言词语相似度。

1 维基百科页面分析

维基百科(Wikipedia)是一个“自由”、“免费”的网络百科全书,是全世界最大的基于 Wiki 技术

的多语种、开放式的在线百科全书。维基百科整个项目总共收录了超过 3 000 万篇概念，具有不同语言的概念对应关系，能通过“跨 Wiki 链接”实现不同语言概念切换[12]。截止到目前共有 285 种语言版本，中文有 856 767 个概念，越南语有1 142 148个概念，这些概念大都具有语言之间对应关系。

维基百科页面结构由左侧“导航”与右侧概念文本构成。左侧“导航”主要为“其他语言”链接，页面右侧是与概念相关的文本描述。例如中文维基百科中“长城”概念，页面右侧为对应的不同语言“长城”概念描述，如越南语“Vạn Lý Trường Thành(长城)”的概念页。在维基百科右侧概念页面，每个概念都是通过文本进行描述的，文本中包含大量词语、实体及短语。利用汉越双语对应的概念页面之间关系，可以构建双语概念特征空间，利用词语在概念页面上的词频及在其他概念页面上与概念的共现关系，构建词语表征的特征向量，实现不同语种的词语之间的相似度的计算。

2　词语的表征

2.1　词向量的构成

维基百科中有多种语言选项，其中中文与越南语概念是计算中-越词语词义相似度的基础，因此，本文选取维基百科中通过“其他语言”列表链接存在翻译对应关系的中文越南语概念，构建一个中文与越南语同时映射的特征空间。定义中文概念集合为 P，越南语概念为 Q。取 $P\cap Q$ 得出一个新的集合 $C=\{c_1,c_2\cdots c_n\}$，即存在翻译对应关系的中文越南文概念合集。如果将词语 A 映射到概念集 C 中，则 A 映射到概念集 C 中的词向量表示为

$$\boldsymbol{a}=\{\omega_{c_1},\omega_{c_2}\cdots\omega_{c_n}\} \tag{1}$$

式中：ω_{c_i}表示词 A 映射在概念集 C 第 i 维上的值，即词语 A 在概念 c_i 上的词频特征值与词语 A 同 c_i 在其他概念文本中共现关系特征值之和。

2.2　词语在概念文本上的词频特征值计算

维基百科页面的概念文本内容是词语映射的基础，直接影响词语在概念中的权重。词语在概念文本中出现的次数与其对于此概念的重要程度成正比增加，同时会因它在获得的所有概念语料集中出现的次数越多，该文越可以认为这个词并不重要。为此，本文定义词语 A 在第 i 个概念 c_i 上的词频特征值 s_i 的计算公式为

$$s_i=tf_i\times\ln\frac{M}{|c_i|} \tag{2}$$

式中：tf_i 表示词语 A 在概念 c_i 中的 tf 值；M 为在所有维基百科概念中收集到的词语总数；$|c_i|$为概念 c_i 文本中的词语数。

考虑到概念页面中的目录类别信息，以及词语出现在概念中位置的不同，词语 A 到根类别的距离 d_i对计算 s_i 帮助。因此，在式(3) 中加入 d_i 进行计算。公式定义为

$$s_i=tf_i\times\frac{\ln(M/|c_i|)}{d_i} \tag{3}$$

2.3　词语与概念的共现特征值计算

维基百科页面中，词语与所在概念页描述的概念在其他概念页面中可能存在共现关系，共现关系次数越大，说明词语与所在概念页的概念越相关。为了有效利用词语与概念在其他概念页面上的共现关系，计算词语 A 与概念共现特征值时，本文考虑了词与概念 c_i 成对出现次数 t 及共现时的最小距离 $\min(l_x)$。当词 A 与概念 c_i 之间距离 l_x 越近，或者出现的次数 t 越多时，两者共现特征值越大。设窗口长度为段落长度 L_x，当它们共现在文本中同一段落时，计算两者共现权重值 k_x，k_x 定义如下

$$k_x=\frac{t\cdot L_x}{|c_x|\min(l_x)} \tag{4}$$

结合词语的词频特征与共现特征对词语 A 映射到第 i 维向量时的权重值的影响，ω_{c_i}表示为

$$\omega_{c_i}=\alpha s_i+\beta\sum_{i=1}^{|x|}k_x \tag{5}$$

式中：α 为词频特征的权重系数，β 为共现特征的权重系数。

3　词语语义相似度计算

词语通过特征空间的映射表征成向量，因此通过计算表征词语的向量之间的余弦值判断两个词语之间的相似度。假设输入中文词 A 与越南文词语 B，两词的向量分别表征为 $\boldsymbol{a}=\{\omega_{a_1},\omega_{a_2}\cdots\omega_{a_n}\}$ 与 $\boldsymbol{b}=\{\omega_{b_1},\omega_{b_2}\cdots\omega_{b_n}\}$，定义词语语义相似度计算公式如下

$$\mathrm{sim}(a,b)=\frac{\sum_{i=1}^{n}(\omega_{a_i}\cdot\omega_{b_i})}{\sqrt{\sum_{i=1}^{n}(\omega_{a_i})^2}\cdot\sqrt{\sum_{i=1}^{n}(\omega_{b_i})^2}} \tag{6}$$

4 实验与分析

4.1 测试集构建

为了验证本文提出方法的有效性，选取了 R&G、M&C 和 WS-353 3 个测试集进行测试。R&G 测试集由 Rubenstein 和 Goodenough 构建，其中包含 65 对词语，旨在研究文本相似度与同义词相似度之间的关系[13]。M&C 由 Miller 和 Charles 从 R&G 测试集中筛选出 30 对词语，并对其进行重新的人工评分构建了一个新的测试集 M&C[14]。WS-353 测试集全称 Wordsimllarity-353 测试集，是 Finkelstein 和 Gabrilovich 在 2001 年创建的。WS-353 测试集包含 353 个词语对，是 3 个测试集中规模最大的一个，它涵盖了 M&C 中的 30 个词语对和 R&G 中大部分词语对[15]。

本文通过翻译 R&G、M&C 和 WS-353 得到汉语和越南语环境下的测试集，这样就可以针对本文提出的汉越双语的词语语义相关度计算方法进行评测。如表 1 所示为 WS-353 测试集，前两列是词语对，第三列是根据评价者的评分得出的这个词语对最终的相关度评分。

表 1 WS-353 测试集示例

Word1	Word2	Human(mean)
love	sex	6.77
tiger	cat	7.35
tiger	tiger	10
book	paper	7.46
computer	keyboard	7.62
computer	Internet	7.58
plane	car	5.77
train	car	6.31
telephone	communication	7.5
television	radio	6.77
media	radio	7.42
bread	butter	6.19

本文工作基于汉语和越南语环境下，需要构建汉语和越南语环境下的测试集。本文通过翻译和筛选原有的英文测试集 WS-353 得到汉语和越南语环境下的包含 200 组词对的测试集。表 2-4 分别为筛选翻译过后的中-中、中-越、越-越测试集。

表 2 WS-353 翻译后中-中测试集示例

Word1	Word2	Human(mean)
爱情(love)	性(sex)	6.77
老虎(tiger)	猫(cat)	7.35
老虎(tiger)	老虎(tiger)	10
书(book)	纸(paper)	7.46
电脑(computer)	键盘(keyboard)	7.62
电脑(computer)	互联网(Internet)	7.58
飞机(plane)	汽车(car)	5.77
火车(train)	汽车(car)	6.31
电话(telephone)	交流(communication)	7.5
电视(television)	广播(radio)	6.77
媒体(media)	广播(radio)	7.42
面包(bread)	黄油(butter)	6.19

表 3 WS-353 翻译后中-越测试集示例

Word1	Word2	Human(mean)
爱情(love)	tình d ục(sex)	6.77
老虎(tiger)	Mèo(cat)	7.35
老虎(tiger)	Hổ(tiger)	10
书(book)	Giấy(paper)	7.46
电脑 (computer)	Bàn phím máy tính (keyboard)	7.62
电脑(computer)	Internet	7.58
飞机(plane)	Ô tô(car)	5.77
火车(train)	Ô tô(car)	6.31
电话 (telephone)	giao lưu (communication)	7.5
电视(television)	đài mồm(radio)	6.77
媒体(media)	đài mồm(radio)	7.42
面包(bread)	Bơ(butter)	6.19

表 4 WS-353 翻译后越-越测试集示例

Word1	Word2	Human(mean)
Tình yêu(love)	tình dục(sex)	6.77
Hổ(tiger)	Mèo(cat)	7.35
Hổ(tiger)	Hổ(tiger)	10
Sách(book)	Giấy(paper)	7.46
Máy tính (computer)	Bàn phím máy tính (keyboard)	7.62
Máy tính(computer)	Internet	7.58
Máy bay(plane)	Ô tô(car)	5.77
Tàu hỏa(train)	Ô tô(car)	6.31
Diện thoại (telephone)	giao lưu (communication)	7.5
Truyền hình (television)	đài mồm (radio)	6.77
phương tiện truyền thông (media)	đài mồm (radio)	7.42
bánh mì(bread)	Bơ(butter)	6.19

4.2　实验数据

实验使用维基百科中中文与越南文语言环境下的概念。爬取概念可以通过左侧“其他语言列表”中“中文”“越南文”存在互译对应关系的概念页面。随机爬取 10 000 对这样的概念文本，构成实验时表征词语的特征空间。实验使用中科院分词工具 ICTCLAS 对概念文本进行分词，再利用停用词列表去停用词。

4.3　评价方法

本文提出的方法计算出的词语相似度与测试集中人工标注结果的一致程度可使用斯皮尔曼等级相关系数(Spearman's rank correlation coefficient)和皮尔森积差相关系数(Pearson product-moment correlation coefficient)衡量。

斯皮尔曼等级相关系数由 Spearman 根据积差相关的概念推导而来，是积差相关的特殊形式。它是一个非参数性质的秩统计参数，用来衡量两个变量之间关联关系的大小。斯皮尔曼等级相关系数用 ρ_s 表示，由式(7)计算得出

$$\rho_s = 1 - \frac{6\sum d_i^2}{n^3 - n} \tag{7}$$

式中：n 代表等级个数，即测试集中包含的词语对的数量；d 代表二列成对变量的等级差数；d_i 表示第 i 个元素的等级差，即算法对第 i 个词语对的相关度评价结果和人工标注结果在各自的排序列表中排序位置的差。

皮尔森积差相关系数是由 Karl Pearson 提出的另一个度量两个变量间相关程度的方法，通常用 ρ 表示。它是一个介于 1 和 -1 之间的值。变量 X 和 Y 的皮尔森积差相关系数是 X 和 Y 的协方差与二者标准差积的商，其计算方法如式(8)所示

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(X-\mu_X)(X-\mu_Y)}{\sigma_X \sigma_Y} \tag{8}$$

4.4　实验结果与分析

为了验证提出方法的有效性，该文作者设计了 2 组实验，通过两种不同评价标准进行评价。

(1) 实验 1 只考虑词语在概念页面上的词频特征计算词语相似度，将通过高维向量表征的英-英、中-中、越-越、中-越的词对分别计算其相似度并进行排序，通过与测试集人工标注相似度排序进行比较。分别得出两种评价标准下的相关系数，见表 5、6。

表 5　仅考虑词语在概念页面词频特征的斯皮尔曼等级相关系数

	En-En	Ch-Ch	Vi-Vi	Ch-Vi
R&G	0.51	0.46	0.47	0.41
M&C	0.72	0.50	0.49	0.46
WS-353	0.70	0.48	0.43	0.45

表 6　仅考虑词语在概念页面词频特征的皮尔森积差相关系数

	En-En	Ch-Ch	Vi-Vi	Ch-Vi
R&G	0.52	0.43	0.38	0.33
M&C	0.57	0.46	0.43	0.42
WS-353	0.55	0.47	0.41	0.40

(2) 在实验 1 的基础上，表征词语时加入词语与概念在其他概念上的共现特征。进行相似度计算。在两种评价标准下得出的相关系数见表 7、8。其中权重系数 α 与 β 分别取 0.3 和 0.7。

表 7　考虑词频特征与概念页面共现特征时的斯皮尔曼等级相关系数

	En-En	Ch-Ch	Vi-Vi	Ch-Vi
R&G	0.53	0.49	0.48	0.44
M&C	0.73	0.54	0.51	0.48
WS-353	0.71	0.51	0.45	0.47

表 8　考虑词频特征与概念页面共现特征时的皮尔森积差相关系数

	En-En	Ch-Ch	Vi-Vi	Ch-Vi
R&G	0.54	0.46	0.42	0.36
M&C	0.59	0.48	0.45	0.47
WS-353	0.57	0.51	0.43	0.42

表 5、7 的数据显示，在中文和越南语各自环境下的对比试验，斯皮尔曼等级相关系数没有在英语环境下高。分析原因，可能由于在人工翻译测试集的时候，出现维基百科中概念的消岐问题，比如说“虎”和“老虎”在维基百科中通过消岐义指向同一个页面，但是在计算的过程中可能会出现映射不到的情况。

表 5、6 与表 7、8 对比发现，同时考虑词频特征与概念页面共现特征时，斯皮尔曼等级相关系数与皮尔森积差相关系数都高于仅考虑词语在概念页面词频特征时，本文中所用方法在汉越词语语义相关度上达到了很好的效果，该方法对汉越词语语义相关度计算是可行的。

5 结束语

借助与维基百科中不同语言概念的描述统计信息,利用词语在概念页面上的统计信息及词语与其他页面共现信息特征,实现不同语言之间词语之间的相关性的计算,结果也证明了提出方法的有效性,利用维基百科资源能够很好实现不同语言词语相关性计算。进一步的研究还可以深入融合更多页面的统计特性,如考虑概念中的链接指出、指入的关系、概念之间的层次关系、概念页面的词语消歧信息等,另外,还将研究如何利用词语相关性分析来计算中-越文本的相似度分析。

参考文献:

[1] Ahsaee M G, Naghibzadeh M, Naeini S E Y. Semantic similarity assessment of words using weighted WordNet [J]. International Journal of Machine Learning and Cybernetics, 2014, 5(3): 479-490.

[2] 田久乐,赵蔚. 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报(信息科学版), 2010, 28(6): 602-608.
Tian Jiule, Zhao Wei. Words similarity algorithm based on tongyici cilin in semantic web adaptive learning system [J]. Journal of Jilin Unversity(Information Science Edition), 2010, 28(6): 602-608.

[3] 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.
Liu Qun, Li Sujian. Word semantic similarity computation based on HowNet [J]. Chinese Computation Linguistics, 2002, 7(2): 59-76.

[4] 詹志建,梁丽娜,杨小平. 基于百度百科的词语相似度计算 [J]. 计算机科学, 2013, 40(6): 199-202.
Zhan zhiJian, Liang Lina, Yang Xiaoping. Word similarity measurement based on BaiduBaike [J]. Computer Science, 2013, 40(6): 199-202.

[5] 张冰怡,魏 博,陈建成,等. 基于对偶编码的中文分词算法 [J]. 南京理工大学学报, 2014, 38(4): 526-530.
Zhang Bingyi, Wei Bo, Chen Jiancheng, et al. Chinese word segmentation algorithm based on pair coding [J]. Journal of Nanjing University of Science and Technology, 2014, 38(4): 526-530.

[6] 王文,王树锋,李洪华. 基于文本语义和表情倾向的微博情感分析方法 [J]. 南京理工大学学报, 2014, 38(6): 733-738.
Wang Wen, Wang Shufeng, Li Honghua. Micro-blogging sentiment analysis method based on text semantics and expression tendentiousness [J]. Journal of Nanjing University of Science and Technology, 2014, 38(6): 733-738.

[7] Dagan I, Lee L, Pereira F C N. Similarity-based models of word cooccurrence probabilities [J]. Machine Learning, 1999, 34(1-3): 43-69.

[8] Gracia J, Mena E. Web-based measure of semantic relatedness [J]. Lecture Notes in Computer Science, 2008, 5175: 136-150.

[9] 赵军,胡栓柱,樊兴华. 一种新的词语相似度计算方法 [J]. 重庆邮电大学学报(自然科学版), 2009, 21(4): 528-532.
Zhao Jun, Hu Shuanzhu, Fan Xinghua. Word similarity computation based on word link distribution [J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2009, 21(4): 528-53

[10] 吴思颖,吴扬扬. 基于中文 WordNet 的中英文词语相似度计算 [J]. 郑州大学学报(理学版), 2010, 42(2): 66-69.
Wu Siying, Wu Yangyang. Chinese and English word similarity measure based on chinese WordNet [J]. Journal of Zhengzhou University(Natural Science Edition), 2010, 42(2): 66-69.

[11] Vulic I, Moens M F. Cross-lingual semantic similarity of words as the similarity of their semantic word responses [C] //Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013). Atlanta: ACL, 2013: 106-116.

[12] 中文维基百科 [EB/OL]. http://zh.wikipedia.org/wiki/Wikipedia, 2016-12-29.
Chinese Wikipedia [EB/OL]. http://zh.wikipedia.org/wiki/Wikipedia, 2016-12-29.

[13] Rubenstein H, Goodenough J B. Contextual correlates of synonymy [J]. Communications of the Acm, 1965, 8(10): 627-633.

[14] Miller G A, Charles W G. Contextual correlates of semantic similarity [J]. Language and Cognitive Processes, 1991, 6(1): 1-28.

[15] Finkelstein L, Gabrilovich E, Matias Y, et al. Placing search in context: the concept revisited [J]. Acm Transactions on Information Systems, 2002, 20(1): 116-131.