基于中文维基百科的词语语义相关度计算[1]

万富强，吴云芳

(北京大学计算语言学教育部重点实验室，北京，100871)

摘 要：语义相关度计算在信息检索、词义消歧、自动文摘、拼写校正等自然语言处理中均扮演着重要的角色。本文采用基于维基百科的显性语义分析方法计算汉语词语之间的语义相关度。基于中文维基百科，将词表示为带权重的概念向量，进而将词之间相关度的计算转化为相应的概念向量的比较。进一步，引入页面的先验概率，利用维基百科页面之间的链接信息对概念向量各分量的值进行修正。实验结果表明，使用本文的方法计算汉语语义相关度，与人工标注标准的斯皮尔曼等级相关系数可以达到 0.52，显著改善了相关度计算的结果。

关键词：语义相关度；显性语义分析；中文维基百科；先验概率；概念向量

中图分类号：TP391　　**文献标识码**：A

Computing Lexical Semantic Relatedness with Chinese Wikipedia

Fuqiang Wan, Yunfang Wu

(Key Laboratory of computational linguistics (Peking University), Ministry of Education, Beijing, 100871)

Abstract: Lexical semantic relatedness plays an important role in natural language processing, such as information retrieval, word sense disambiguation and automatic text summarization and spelling correction, etc. In this paper, we employ Wikipedia-based Explicit Semantic Analysis to compute semantic relatedness between Chinese words. Based on Chinese Wikipedia, a word is represented as weighted vectors of concepts. Then, computing the semantic relatedness of words amounts to comparing the corresponding concept vectors. Furthermore, we add the priori probability factor of concept and use the linking information among the Wikipedia pages to optimize the concept vectors. The experimental results show that the Spearman's rank correlation coefficient between the computed relatedness and human judgments reaches 0.52, which significantly outperforms the baseline.

Key words: semantic relatedness; explicit semantic analysis; Chinese Wikipedia; priori probability; concept vectors

1　引言

语义相关度的计算在很多自然语言处理的应用中都扮演着重要的角色。信息检索[1]系统中使用相关度得分，对查询进行扩展。词义消歧[2]一直以来都是计算语言中一个比较难解的问题。利用词语之间的相关性能够协助计算机进行词义消歧。例如，“削苹果的刀”与“削苹果的皮”，两者都是“动词+名词+助词+名词”的结构，可以利用“苹果”与“刀”，“苹果”与“皮”的相关度对两者加以区分。此外，在文档自动文摘以及问答系统中常常使用相关度或相似度的得分，评估候选语句的精准程度。在拼写校正[3]中也会用到语义相关度的计算。研究如何更好地计算文本或者词汇之间的语义相关度是一个重要的课题。

本文研究基于 Gabrilovich & Markovitch 提出的基于维基百科的显性语义分析（Explicit Semantic Analysis, ESA）方法[4]，对中文词语之间的语义相关度进行计算。将词表示为带权重的概念向量，计算目标词语之间的相关性就转化为比较相应的概念向量。本研究选取的概念由中文维基百科文章明确定义，即将指定的中文维基百科的一个页面作为一个概念，引入概念(页面)的先验概率，利用维基百科词条的词频信息和页面之间的链接信息对算法进行了多种改进。实验结果表明，引入页面先验概率因子，能够明显改善目标词对相关度计算的结

[1] 基金项目：教育部人文社会科学研究规划基金项目(13YJA740060)，国家社科基金重大项目(12&ZD227).

果—斯皮尔曼等级相关系数从 0.40 提高到 0.52。

本文组织结构如下：第 2 节介绍了前人的相关工作；第 3 节阐述显性语义分析方法的核心思想及引入页面先验概率的改进算法。第 4 节介绍中文维基百科概念的选取，实验采用的评测数据集以及评测的指标—斯皮尔曼等级相关系数。第 5 节展示本实验的结果及对结果的分析。本文最后，即第 6 节对本实验进行了总结。

2 相关工作

语义相关度的计算可以划分为 3 类方法：基于大规模语料库的方法、基于语义分类体系的方法和基于百科知识的方法。基于大规模语料库计算文本（或单词）的相似度或者相关度，主要有两种方法: 一种方法是简单地使用词语共现信息。该方法假定同时出现在文档或者段落中的词在某种意义上相似或者相关，它将文档或者段落视为词的集合，忽略词与词之间的语法信息。另一种方法是对文档或者段落进行浅层的句法分析，得到词汇之间语法关系或者依存关系，在依存分析结果的基础上进行相似度计算。使用词共现信息更具有鲁棒性，不会涉及语句的句法分析，实现起来更加简单。目前有许多关于语义相关度和相似度的研究是基于前一种方法的[17,18,19]。

英语中基于语义分类体系计算语义相关度主要是依据 WordNet[5]，而汉语中主要是依据 HowNet。前人基于 WordNet 的层次分类体系实现的词汇语义相似度度量方法有以下 4 种：1). 边计数方法。如果该网络中的两个概念 c_1，c_2 之间的连接越多，两个概念之间的距离越短，那么它们就越相似。具体度量方法有：最短路径[6]，带权重的链接[7]等。2). 信息含量方法。两个概念的相似度与它们共享的信息相关，而共享信息是由在网络层次体系中涵括它们的高层的概念表征。如 Resnik[8]，Lin[9]等工作。3). 基于特征的度量方法。每一个词都由能表征它性质、特征的词的集合表示，如 Tversky[10]。4). 组合方法，如 Rodriguez et al.[11]。

随着维基百科的普及和盛行，近年来出现了一些基于百科知识的相关度计算方法。Michael Strube 等提出使用 Wikirelate!方法[12]计算词语之间的语义相关度，该方法首先将两个目标词 t_1，t_2 用以它们为标题的文章来表示，并提取文章的类别信息，然后使用基于文本覆盖的方法，或者利用维基百科的类别树，使用基于路径或信息含量的方法计算两篇文章的相关度，也即是两个目标词的相关度。Gabrilovich and Markovitch 提出基于维基百科的显性语义分析方法（Explicit Semantic Analysis，ESA）[4]用于计算文本（或词）之间的语义相关度。孙琛琛等[18]利用英文维基百科结构信息计算语义关联度。李赟等[19]利用中文维基百科进行语义相关词的获取及其相关度分析。

还有研究者利用其他的资源进行语义相关性研究。如利用维基词典计算语义相关性[13]，使用网络搜索引擎度量词语之间的相似度[14]等。Torsten et al.[15]的研究表明，基于 German WordNet 的语义相似度度量方法比基于维基百科的语义相似度度量方法更接近人工判定的结果；然而，基于维基百科的语义相关度度量却比基于 German WordNet 的语义相关性度量方法要好。

3 语义相关度计算的基本方法

3.1 基本原理

分布相似在一定程度上能够反映语义相似以及语义相关，因此可以将词语之间的语义相关性度量转化为词语分布的相似性度量。显性语义分析(ESA)，是将词表示为带权重的概念向量，计算词语之间的相关性就转化为比较相应的概念向量。本文选取的概念由中文维基百科文章明确定义，即将中文维基百科的页面作为概念。

令 N 表示中文维基百科的单词数(即词汇表 L 的大小)，M 表示选取的概念(页面)数。用 $w_{i,j}$ 表示词项 t_i 与概念 c_j 的关联程度。该值越大，表明词 t_i 与该概念 c_j 的关联程度越强；反

之，则表明词 t_i 与该概念 c_j 的关联程度越弱。词-文档矩阵表示为：

$$P=\begin{pmatrix} w_{1,1} & \cdot & \cdot & w_{1,M} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ w_{N,1} & \cdot & \cdot & w_{N,M} \end{pmatrix} \tag{1}$$

则词 t 的概念向量 V 可表示为：

$$V=\begin{cases} 0, \text{若t 不在词汇表L中} \\ (w_{k,1}, \ldots, w_{k,M}), t\text{是词汇表L中的第k个词项} \end{cases} \tag{2}$$

然后，根据概念向量 V_1 和 V_2，使用 cosine 方法比较两个向量，计算目标词对<t_1, t_2>的相关度（当至少有一个目标词不在词表中时两者的相关度记为 0）：

$$rel(t_1,t_2)=\frac{V_1\cdot V_2}{|V_1||V_2|} \tag{3}$$

3.2 基本 TFIDF 方法

Gabrilovich 等提出的基于维基百科的 ESA 方法[4]采用在信息检索中常用的 TFIDF（即词项频率与逆文档频率的乘积）作为词与文档的关联程度的度量。使用数学公式表示为：

$$w_{t,c}=TF_{t,c}\times IDF_t=TF_{t,c}\times\log(M/DF_t) \tag{4}$$

由于 IDF_t 仅由词 t 决定，对于同一个 t 而言 IDF_t 是相同的。使用余弦相似度方法比较词的概念向量时，对向量长度进行了归一化，因此事实上 IDF_t 并没有真正参与到计算之中，结果仅由 $TF_{t,c}$ 决定。于是，可以将各个分量都含有的常量提出来，记为 k。目标词 t 的概念向量可以简单的表示为：

$$V=k(TF_{t,1},\ldots,TF_{t,M}),\text{其中}k=\log(M/DF_t) \tag{5}$$

为了便于表述，将此方法记作 TFIDF。

4 语义相关度计算的改进方法

利用显性语义分析（ESA）方法，使用 TFIDF 作为权值度量，计算汉语语义相关度的结果并不理想。本文引入页面的先验概率，提出了以下的改进方法。

在信息检索中使用查询似然模型，将文档按照其与查询相关的似然 P(d|q)排序。查询似然模型是信息检索中最早使用也是最基本的语言模型。P(d|q)度量了 d 与 q 的相关性程度。利用贝叶斯公式有 P(d|q)=P(q|d)P(d)/P(q)。将词 t 与 q 对应，概念 c 与 d 对应，我们得到词项 t 与 c 关联程度：

$$w_{t,c}=P(c|t)=\frac{P(t|c)P(c)}{P(t)} \tag{6}$$

对 P(t|c)使用最大似然估计，有：

$$\hat{P}(t|c)=\frac{TF_{t,c}}{\sum_{t'\in L}TF_{t',c}}=\frac{TF_{t,c}}{T_c} \tag{7}$$

对于给定的 t，P(t)是一个常数，于是有：

$$w_{t,c} \propto \frac{TF_{t,c} \times P(c)}{T_c} \tag{8}$$

$$V = k(TF_{t,1} \times P(c_1) / T_{c_1}, ..., TF_{t,M} \times P(c_M) / T_{c_M}), \text{其中} k = 1 / P(t) \tag{9}$$

$TF_{t,c}$ 以及 T_c 通过对中文维基百科数据进行分词以及词频统计便可得到，因此为了得到词 t 与概念 c 的相关程度 $w_{t,c}$，只需对先验概率 P(c)进行估计。比较公式(5)和公式(9)，基本的 TFIDF 方法，等价于取 c 的先验概率正比于词条数目的模型。然而，仅使用文档词条数目作为文档先验概率的估计因子有失偏颇，本文提出以下方法对页面(概念)的先验概率进行估计：

4.1 NORM_TF 方法

对 P(c)进行估计最简单的方法便是，所有概念 c 出现的概率相同。即对于任意的 c，P(c)是一个定值（此处取为 1/M）。同样由于使用 cosine 方法比较词与词的概念向量，因此，词项 t 的概念向量 V 可以简单记为：

$$V = k(TF_{t,1} / T_{c_1}, ..., TF_{t,M} / T_{c_M}), \text{其中} k = 1 / P(t)M \tag{10}$$

该向量与 TFIDF 基本方法得到的概念向量差别在于，它对词项频率（TF）进行了归一化。为了表述的方便，将此方法记为 NORM_TF。

4.2 INLK 方法

前文提及在进行 Wikiprep 处理的同时得到了页面之间的链接信息。维基百科页面之间的链接与普通网页链接有所不同。普通网页链出的数目较少，而维基百科页面的链出很多。维基百科的链接是这样生成的：如果在一个页面中出现了某个词（或词组），而这个词（或词组）正好又是维基百科的一个词条，那么该页面就有一条指向词条对应页面的链接。如页面“阿波罗计划”中出现了词‘苏联’，而锚文本“苏联”又正好是维基百科的一个词条，对应了维基百科的一个页面，因此从页面“阿波罗计划”到页面“苏联”有一条链接。

由于维基百科页面的链接信息在一定程度上能够反映页面被访问的频率。考虑到维基百科链接构造的特殊性，可以认为越频繁出现的词条，其对应页面的入度越大，页面被访问的频率越高。基于这个假设，记页面（概念）c 的入度为 $INLK_c$，则可以对 P(c)进行估计。由于选取的概念入度差别非常大，因此直接使用入度进行计算会使得页面入度大的 P(c)非常大，因此可以对入度采用取对数的方法，此时概念 c 的先验概率 P(c)表示为：

$$\hat{P}(c) = (\log(INLK_c) + 1) / \sum_{c' \in C} (\log(INLK_{c'}) + 1) \tag{11}$$

同样，为了便于表述，将此方法记为 INLK.

4.3 PRANK 方法

既然提及页面之间的链接，自然就会想到 PageRank[16]。记网页数量为 K，根据 Web 图的邻接矩阵 A（K×K），并记 A 第 i 行 1 的个数为 N_i，可以推导出该马尔科夫链的概率转移矩阵 P（K×K）：

$$P(i;j) = \begin{cases} 1/K, \text{如果} N_i = 0 \\ (1-\alpha)/N_i + \alpha/K, \text{如果} A(i;j)=1 \\ \alpha/K, \text{其他} \end{cases} \tag{12}$$

初始化访问各网页的概率为：$\vec{\pi} = (1/K, ..., 1/K)$，不断的使用矩阵 P 右乘 $\vec{\pi}$，迭代一定次数之后会收敛于一个稳定的值，此时得到的向量 $\vec{\pi} = (v_1, ..., v_K)$ 作为最终各页面访问频率的向量。

对中文维基百科的概念使用上述方法（取 α =0.1），可以得到各个概念被访问的频率，使用它对 P(c)进行估计。与 INLK 方法一样 v_c 的差距很大，但不能像 INLK 方法那样先取对数再加 1，因为直接取对数得到的是负值。于是将 v_c 乘以 10M（M 为选取的概念的个数），使得其值大于等于 1。再对该结果取对数加 1。P(c)的估计值为（将此方法记为 PRANK）：

$$\hat{P}(c) = (\log(10Mv_c) + 1) / \sum_{c' \in C} (\log(10Mv_{c'}) + 1) \quad (13)$$

4.4 TDF 方法

维基百科词条有着对其页面内容的充分概括性，页面内容都是对该词条的阐述。因此可以使用页面的标题在整个数据集中出现的频率（CF）或者文档频率（DF）来度量概念的先验概率 P(c)，使用 TCF 表示概念标题（词条）的 CF，TDF 表示概念标题的 DF，并采用对数平滑方法，则对 P(c)的估计分别为：

$$\hat{P}(c) = (\log(TCF_c) + 1) / \sum_{c' \in C} (\log(TCF_{c'}) + 1) \quad (14)$$

$$\hat{P}(c) = (\log(TDF_c) + 1) / \sum_{c' \in C} (\log(TDF_{c'}) + 1) \quad (15)$$

同样为了表述方便，将两种对估计 P(c)计算词与词之间相似度的方法分别记为 TCF，TDF。

4.5 COMB 方法

前文已使用了多种方法对 P(c)的值进行估计，如 INLK，TDF 等。考虑到他们的组合实在是太多，但都是基于维基百科链接或者维基百科页面的标题，因此仅仅选取他们两两组合中的其中一种，即 TDF+PRANK（记为 COMB），前者基于标题词频，后者基于链接，并且使用最简单的线性组合的方式将两者对概念的先验概率的估计加以组合，即：

$$w_{t,c} = \alpha w_{PRANK(t,c)} + (1-\alpha) w_{TDF(t,c)} \quad (16)$$

其中，$w_{PRANK(t,c)}$以及$w_{TDF(t,c)}$ 分别表示使用 PRANK 和 TDF 方法得到的权重。

5 评测

5.1 概念选取

从中文维基百科网站(http://zh.wikipedia.org/)下载中文版维基百科的 XML 转储数据(zhwiki-20101029-pages-meta-current.xml.bz2)，数据解压后使用 Wikiprep[2]处理，去掉模板页面、重定向页面、类别页面等以及页面中无关的域（仅保留页面标题、页面 ID 以及文本域）。进行 Wikiprep 处理的同时会得到页面的链接信息以及类别信息等。由于中文维基百科页面中包含简体和繁体中文，我们使用中文繁简转换工具，统一将所有的繁体字转换为简体字。得到 1G 的文本文件，共有 324216 个页面。

有些中文维基百科页面的正文太短，包含的信息量很少，编辑的内容质量不高。如果将

[2] 从 http://search.cpan.org/~triddle/Parse-MediaWikiDump-1.0.4 下载。原始代码用于处理英文维基百科数据，修改部分代码之后即可用于处理中文维基百科的数据。

所有的页面都作为最终的概念，那么得到的词的概念向量的维度很大，在很多维度上噪音很大，对词相关度的计算造成不利的影响。因此需要在这些页面中选出一个子集 C 作为最终概念集合。由于页面入度和词数在一定程度上能够反映页面的质量，因此在实验中去掉了入度过小（小于 3）或者词数过少（少于 70）的页面，剩下的页面（127936 个）即作为最终的概念集合 C，用于词语之间相关度的计算。

为了统计页面的词条数，本实验使用了中文停用词表[3] ，对概念集合 C 中所有的页面进行自动分词。维基百科页面标题通常是人名、地名、专有名词等，因此为了将它们作为一个词（或词组）保留下来，实验时将页面标题作为一个词条。由于这些词条数目众多，不可能人工对其进行词性标注，而缺少词性标注会对分词结果造成影响。为了降低这种不良影响，采取了以下措施：首先使用中文分词器[4]对这些词条进行分词，将分词器不能识别的词条（分词器会将其切分开）加入到用户词典，再次使用分词器对维基百科数据进行分词。对概念集合进行解析，统计词条（token）数目 T_c 的同时，得到了以下数据：(1)词汇表 L;(2)词项 t 在多少个概念中出现 DF_t;(3)词项 t 在概念 c 中出现的频次 $TF_{t,c}$;(4)词项 t 在所有概念中出现的频次 CF_t。

5.2 评测数据

本实验的评测数据基于英文 WordSimilarity-353 数据集，这是英语语义相似度研究中广泛应用的一个评测标准。根据 WordSimilarity-353[5]得到中文词相关度测试的数据集（为了便于表述，将此数据集记为 ZH-SIM-353），具体做法如下。

首先，两个计算语言学研究生独立地对数据集 WordSimilarity-353 进行人工翻译，将英语单词对翻译为汉语词语对，然后让第三者对前两者翻译不一致的词对进行修改。只有当词对中的两个词都翻译得完全相同时才称为一致。WordSimilarity-353 总共有 353 个词对，其中两人翻译一致的词对数为 169，占总数的 48%。两人翻译不一致时，再进行如下处理：

1.单字词和双字词。两个翻译者在单字词和双字词的使用上显现出差异，如表 1 所示。解决方法：让翻译结果音节一致；不一致时，倾向于双音节。例如在表 1 中，得到的翻译正确结果为{<虎，猫>，<老虎，老虎>，<药物，滥用>}。

表 1：单字词 VS 双字词

英文词对		翻译 1		翻译 2	
word1	word2	词 1	词 2	词 1	词 2
tiger	Cat	虎	猫	老虎	猫
tiger	Tiger	虎	虎	老虎	老虎
drug	abuse	药物	滥用	药	滥用

2. 别名的使用。如 potato 一者翻译为“土豆”，另一者翻译为“马铃薯”。解决方法：使用更通用的称说，此处选择“土豆”作为 potato 的中文翻译。

3. 去掉翻译为汉语时有明显歧义的 5 个词对，它们分别是<stock, egg>，<stock, live>，<brother, monk>，<crane, implement>以及<life, term>。将剩下的 348 个中文词对以及它们的得分，作为最终的评测集。

5.3 评测指标

本实验采用斯皮尔曼等级相关系数对目标词对的相关度计算的结果与人工标注评测集 ZH-SIM-353 的一致性进行评价。.斯皮尔曼等级相关系数是反映两组变量之间联系的密切程度，它和相关系数 r 一样，取值在-1 到+1 之间。斯皮尔曼等级相关系数的计算公式如下：

[3]总共有 1208 个停用词，可从 http://www.hicode.cn/download/view-software-13784.html 下载
[4]中国科学院计算技术研究所开发的 ICTCLAS 汉语分词系统，http://ictclas.org/ictclas_download.aspx 下载
[5] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

$$r_R = 1 - \frac{6\sum_{i=1}^{n}(R_{Xi} - R_{Yi})^2}{n(n^2-1)} \quad (17)$$

其中n为样本容量, R_X为变量X的等级数，R_Y为变量Y的等级数

6 实验结果与分析

6.1 相关性系数

使用各种向量的权值计算方法对目标词对之间的相关度进行计算，然后按照相关度值降序排列得到词对的等级数，其结果与人工判断标准的斯皮尔曼等级相关系数见表 2：

表 2：不同方法的斯皮尔曼等级相关系数

方法	相关系数	方法	相关系数
TFIDF(基准)	0.403	NORM_TF	0.484
INLK	0.500	TCF	0.510
PRANK	0.512	TDF	0.513
COMB	**0.515**		

从表 2 中可以看出，本文提出的改进方法 NORM_TF，INLK，PRANK、TCF，TDF 以及集成方法 COMB 均比基本方法 TFIDF 有显著提高。即对词与词之间相关性的度量与人工判定的结果更一致，在评测集 ZH-SIM-353 上明显优于基本方法——TFIDF 方法。结果表明：明确地引入概念(页面)的先验概率，利用维基百科页面链接信息，修正词向量元素的值可以提高相关度计算的结果。

6.2 概念数量的影响

前文已经提到由于有些页面正文太短，页面的质量可能较低，重要性不够，有些页面的入度很小，即没有指向它的链接或指向它的链接很少，因此在实验中去掉了入度过小或者词数过少的页面，将剩余的页面作为最终的概念。我们探究了作为概念的页面入度的下界 a，以及词数的下界 b 对计算词-词之间的相关度的影响。

为了选择较好的概念集合，采用实验结果较好的 PRANK 方法和 TDF 方法，对参数 a 以及 b 进行调节。不同 a，b 对应不同的概念集合，采用不同的概念集合计算词与词之间的相关度的结果会有所不同，表 3 列出了概念数目以及实验结果的斯皮尔曼等级相关系数随 a，b 变化的情况。为了更好地观察实验结果随 a，b 变化的趋势将上表转化为曲线图，如图 1 所示（其中实线和虚线分别代表采用 PRANK 方法和 TDF 方法对目标词对的相关度计算结果与 ZH-SIM-353 人工标注结果的斯皮尔曼等级相关系数的变化）。

表 3：概念的选取

a	b	concepts	r(PRANK)	r(TDF)
2	60	153535	0.5088	0.5078
3	60	136327	0.5106	0.5120
4	60	121615	0.5106	0.5086
5	60	109451	0.5106	0.5106
2	70	143372	0.5084	0.5057
3	70	127936	0.5118	0.5133
4	70	114756	0.5120	0.5119
5	70	103733	0.5107	0.5126
2	80	134560	0.5096	0.5096

3	80	120667	0.5119	0.5133
4	80	108715	0.5122	0.5120
5	80	98640	0.5115	0.5141

从图 1 可以看出，当 a，b 变化时，目标词对相关性计算的结果也随着变化，但是结果与 ZH-SIM-353 的一致程度并没多大变化，仅仅有细微的波动。因此在一定范围内 a，b 的取值对相关度计算的结果几乎没有影响。TDF 和 PRANK 方法对概念集合的选取具有较强的鲁棒性。

从表 3 可以看出，当<a, b>=<2,50>时概念数量比 a, b 取其他值时多，但是相关度计算的结果却比其他很多时候都差一点，这说明并不是概念的数量越多越好，当然也不是越少越好（从<a, b>=<4, 70>以及<a, b>=<5, 80>可以看出）。在 a，b 变化时，两种方法计算相关性的结果仍然非常接近，可以说明两者在对概念（concept）的先验概率的估计上是比较一致的。这种一致性很大程度上是由中文维基百科页面之间链接的特殊性决定的。

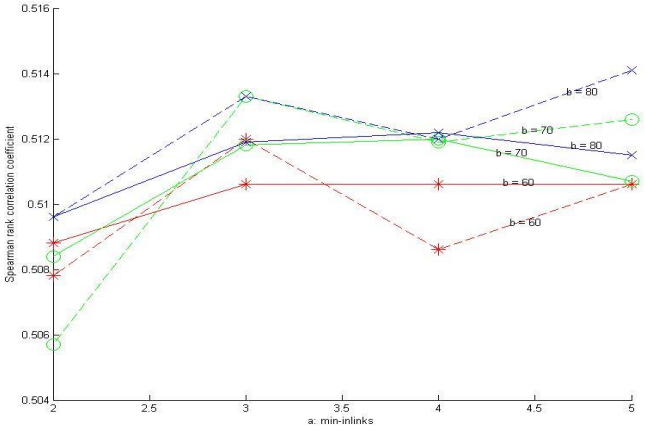

图 1：概念集合对结果的影响

6.3 集成方法中参数的选择

为了探究 COMB 方法中参数 α 的取值变化对词-词相关度计算实验结果的影响，我们针对不同的参数 α，得到目标词对相关度与人工标注的 ZH-SIM-353 数据的斯皮尔曼等级相关系数，见图 2。

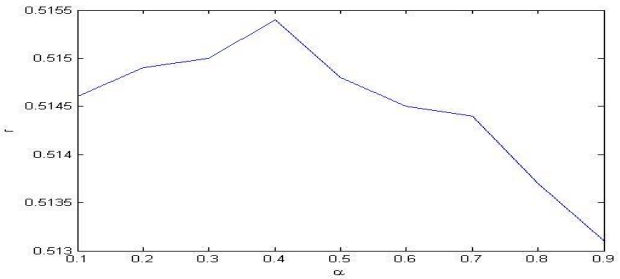

图 2：组合方法参数选取

从图 2 可以看出参数 α 的变化会使得实验结果的斯皮尔曼等级相关系数有些微的变化，当 α 取 0.4 时，在测试集 ZH-SIM-353 上表现得最好。但是随着参数 α 的变化，实验结果并没有显著的变化，斯皮尔曼等级相关系数波动幅度非常小（不到 0.003），这也说明了 TDF 方法和 PRANK 方法对概念 c 的先验概率 P(c)的估计很一致，两种方法计算词与词之间的相关度的结果也比较一致。

7 结语

本文研究采用显性语义分析方法，基于中文维基百科实现了汉语词与词之间的相关度计算。基本方法是，将词表示为带权重的由中文维基百科文章定义的概念向量，将词之间的相关度计算转化为比较相应的概念向量，然后，使用余弦方法比较两个向量，得到词之间的相关度。本文改进方法中，利用概率模型，引入概念的先验概率，利用维基百科文章标题的文档频率、文档集频率以及页面之间的链接结构信息对概念的先验概率进行估计。实验结果表明，本文的改进方法显著提高了相关度计算性能，斯皮尔曼等级相关系数从 0.40 提高到 0.52。文章进一步比较分析了各种方法的特点，并指出在一定范围内，概念集合的选取对词语之间相关度计算结果的影响甚小，组合方法参数的选取对相关度计算的结果也几乎没有影响，我们提出的改进方法具有较强的鲁棒性。

本文研究的测试集是从英文测试集翻译而来。然而，中英文词之间并没有一一对应的关系。为了检验本文提出的改进方法是否与本研究采用的测试集有关，它是否也同样适用于英文，未来的工作有两个方面：其一，在其它的中文相关度测试集上对本文的方法进行测试，观察评测结果是否与本文的结果一致；其二，使用英文维基百科在英文的测试集上检验该改进方法是否同样适合于英文。

参考文献

[1] Finkelstein, L., E. Gabrilovich, Y. Matias, et al. Placing search in context: The concept revisited. ACM Transactions on Information Systems, 2002, 20, 1, 116-131.

[2] Patwardhan, S., S. Banerjee&T. Pedersen. SenseRelate::TargetWord—A generalized framework for word sense disambiguation. Proceeding of AAAI-05, 2005.

[3] Budanitsky, A. & G. Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, 2006, 32, 1, 13-47.

[4] Gabrilovich, E. and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of IJCAI, 2007, pp. 1606-1611.

[5] Fellbaum, Christiane(editor). WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press, 1998.

[6] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man, and Cybernetics, January/ February 1989, 19, 1, 17-30.

[7] R. Richardson, A. Smeaton, and J. Murphy. Using WordNet as a Knowledge Base for Measuring Semantic Similarity BetweenWords. Technical ReportWorking paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland, 1994.

[8] O. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. Journal of Artificial Intelligence Research, 11:95-130, 1999.

[9] D. Lin. Principle-Based Parsing Without Over generation. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93), pages 112-120, Columbus, Ohio, 1993.

[10] A. Tversky. Features of Similarity. Psychological Review, 84(4):327-352, 1977.

[11] M.A. Rodriguez and M.J. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering, 15(2):442-456, March/April 2003.

[12] Michael Strube and Simon Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In AAAI'06, pp.1419-1224, Boston, MA, 2006.

[13] Zesch, T., Müller, C., Gurevych, I.: Using Wiktionary for Computing Semantic Relatedness. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, pp. 861–867 (2008).

[14] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using Web search engines. Proceedings of WWW, 2007.
[15] Torsten Zesch, Iryna Gurevych, and Max Muhlhauser. 2007b. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In Proceedings of NAACL-HLT. Rochester, New York, pages205-208.
[16] L. Page, S. Brin, R. Motwani, and T. Winograd, The PRANK Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper, 199.
[17] 石静, 吴云芳, 邱立坤, 吕学强. 2013. 基于大规模语料库的汉语词义相似度计算方法.《中文信息学报》第 1 期, pp1-6.
[18] 孙琛琛, 申德荣等. 2012. WSR:一种基于维基百科结构信息的语义关联度计算算法.《计算机学报》第 11 期, pp2361-2370.
[19] 李赟, 黄开妍等. 2009. 维基百科的中文语义相关词获取及相关度分析计算.《北京邮电大学学报》第 3 期, pp109-112.