收稿日期：2017-03-20
基金项目：国家自然科学基金项目（201361235636）；湖北省科技支撑计划软科学项目（2013AKB12）
作者简介：张三（1991－），男，武汉大学计算机学院硕士研究生，研究方向为自动推理与符号计算；李某四（1968—），男，博士，北京科技大学计算机与通讯工程学院教授、博士生导师，研究方向为自动推理与符号计算、程序验证。本文通讯作者：李某四。

批注 [A1]: 基金项目来源（编号）

批注 [A2]: 在读学位，标注博士研究生或硕士研究生

批注 [A3]: 若有必要加注通讯作者的，按此格式；通讯作者非单纯意义的“通讯联系人”，担负着文章可靠性的责任。

《软件导刊》论文格式说明

张　三[1]，李某四[2]

（1. 武汉大学 计算机学院，湖北 武汉 430072；2. 北京科技大学 计算机与通讯工程学院，北京 100083）

批注 [A4]: 标注对应作者的工作单位

批注 [A5]: 学校　学院（系），单位所在地（直辖市或省份　城市）邮编

摘　要：【目的】为了改善传统实体解析算法在单机环境下采用人为设定属性权值及阈值难以对海量数据进行快速有效处理的缺点。【方法】基于 Hadoop 框架使用 MapReduce 计算模型，在多节点分布式环境下，通过不断调整网络学习属性之间的内在关系以及属性权值、阈值等参数，将模型放在 Hive 数据仓库中真实数据集上进行有效性验证，分别使用 5000 条数据以及 9000 条数据进行实验。【结果】实验结果表明，基于学习的并行实体解析算法准确率、召回率和 F1 值较高，分别达到了 97.5%、96%和 99。【结论】基于学习的并行实体解析算法不仅能快速有效处理海量数据，而且能有效降低人工经验中存在的误差，同时也能提高识别结果准确度，提升识别效率。

批注 [A6]: 一般论文采用报道式摘要，按“目的、方法、结果、结论”的范式撰写，200-300 字为宜；综述类文章除外。

关键词：数据仓库；实体解析；重复记录；神经网络；时变参数

批注 [A7]: 需是反映论文主题概念、具有检索意义的词或词组，具有专指性，3-6 个为宜。

开放科学（资源服务）标识码（OSID）：

批注 [A8]: 请查看微信公众号或官网通知创建 OSID 码

DOI：
中图分类号：TP3　　文献标识码：A　　文章编号：1672—7800（2017）00

批注 [A9]: 编辑部统一处理

Software Guide Paper Format

批注 [A10]: 英文题名实词首字母大写；超过 5 字符的虚词首字母大写。

ZHANG San[1]，　LI Mou-si[2]

(1. School of Computer Science, Wuhan University, Wuhan 430072, China; 2.School of Computer & Communication Engineering, University of Science & Technology Beijing , Beijing 100083, China)

批注 [A11]: 作者姓全部大写，名的首字母大写；名若为多个字，中间用“-”连接。

批注 [A12]: 作者单位的英文翻译：(学院，学校，城市　邮编，China)

Abstract: For solving the disadvantage in traditional entity resolution algorithm which is usually used in the single machine environment setting the artificial attribute weights and threshold processing methods for entity analysis, which makes the recognition result heavily dependent on manual experience and difficult in efficient big data processing, this article tries to study the intrinsic relationship between the attributes through adjustment network in multiple-nodes-distributed environment by using Map Reduce Calculation Model based on Hadoop Frame. Through adjusting attribute weight and threshold value we can validate on the real data set in the Hive data warehouse by using separately 5,000 and 9,000 data records. Experiment result have shown that parallel entity analysis algorithm based on self-learning has higher accuracy, recall value and F1 value, thus we can draw the conclusion that parallel entity analysis algorithm based on learning has not only effectively reduced the errors in the artificial

experience, which made the recognition result obtain high recognition accuracy and recognition efficiency, but can also deal with the massive data with high efficiency.

Key words: data warehouse; entity resolution; duplicate records; neural network; time-varying parameters

批注 [A13]: 英文关键词小写(专有名词除外），用分号隔开。

0 引言

批注 [A14]: 标题序号按 0；1、1.1、1.1.1……标注，下同

随着面向服务的体系架构（*Services Oriented Architecture,SOA*）的发展，*Web* 服务已经逐渐成为在提高分布式应用程序的灵活性和可扩展性等方面的有效解决方案，被广泛地应用在各个领域，并将成为下一代商业服务应用运行的基石[1]。Web 服务标准和技术发展已经逐步成熟，面对网络上分布的大量功能属性相同，服务质量 *QoS(quality ofservice)*不同的服务,可供用户选择的服务资源越来越多，因此有效的服务选择方法显得尤为重要，已经成为服务计算领域需要解决的核心问题之一。

批注 [A15]: 专业术语有英文缩写的，在**正文中第一次出现时**需写全中文名称、英文及缩写，如：
大规模网络开放课程 (Massive Open Online Courses, MOOC)。

……

对功能属性相同的服务，传统的服务选择方案是针对 *QoS* 各属性值分别赋予不同的权重，进行简单相加，最后将分数最高的服务返回给用户。为加强对服务运行风险因素的评估，人际网络中的信任概念被引入计算机系统。国内外已有部分学者对此进行了研究，并取得了相应成果。 *Maximilien* 等[2]对 *Web* 服务的信誉模型进行了研究,其方法主要依靠用户的反馈,通过用户主观投票打分的统计值来定义信誉度，但它忽略了 *QoS* 属性值的可信性。因为无法保证每个用户使用服务后都具有提供合格反馈等级的能力，若将这些鱼龙混杂的反馈一视同仁的对待，势必会带来评价的偏差，这将直接影响服务选择结果的可信性。陈**等[3]提出……,解决了**问题,其方法具有普适性,但在**方面存在不足。文献[4]提出了……文献[5]分析了……

批注 [A16]: 参考文献实引标注(一)：
1、文中未出现作者姓名时，序号标注于引用内容句末右上角；
2、按[1]、[2]…全文编号，与文后参考文献列表对应；为方便核对，请用黄色突出显示，下同。

批注 [A17]: 参考文献实引标注(二)：
1、文中出现作者姓名时，序号标注于作者姓名右上角；
2、多位作者的，只需写出第一作者名“**”，用“**等”表示；
3、英文人名，只需写出姓氏。

以上方法都未在**方面……本文通过**方法，进行了**研究，改进了/弥补了……。

批注 [A18]: **需有详尽的文献综述，以体现本文的研究意义与创新价值：**
① 介绍该研究领域**近年来（查新）**的**国内、外（查全，一定要有国内文献）**发展和现状，并与本文研究进行充分比较，提出本文的改进之处或创新点；
② 明确标明哪些工作是自己的研究成果，哪些是对他人工作的介绍，在介绍他人工作时，请**标明引证来源**，并作为参考文献列出。

1 词语相似度研究现状

词语相似度主要分为基于语义本体资源、基于统计算法和将前两者融合的混合技术三种方法：利用语义资源计算词语相似度也可称为基于本体（或知识库）的词语相似度算法，主要根据专家人工建立的语义网络计算相似度。

1.1 基于语义资源的词语相似度算法

近年来，一些诸如同义词词林、WordNet、知网这种大规模可量化的语言本体的诞生与发展，为进行真实文本的语义分析和理解提供了强有力的资源支持。特别是最近几年“知网”等语义资源不断丰富发展，中文语义研究方向逐渐增多。知网作为一个知识系统，是一个网而不是树。

1.2 基于统计的语义相似度算法

基于统计的语义相似度方法建立在如果两个词语的含义相同或相近,则伴随它们同时出现的上下文也相同或相近。该方法主要以词语的上下文信息的概率分布作为相似度参考，计算方法主要有向量空间模型（VSM）、词语共现信息、基于部分语法分析和改进的基于大规模语料库的方法。

定义1 有三条边、三个顶点的多边形叫作三角形。

批注 [A19]: 定义、定理、算法、公式、举例等，分别从 1 开始标号，全文排序。

定理1 三角形的内角和等于180度。

算法1 服务器调度算法。

……

1.3 基于混合技术的语义相似度算法

基于大规模语料统计的算法相对专家手工建立的语义资源更加客观，但每种统计模型的创建都受语料库中数据质量的极大干扰……所以，一种语义与统计相融合的词语相似度算法应运而生，通过发挥两种算法各自的优势进行词汇间的语义相似度的计算。

2 词语相似度应用实验

FAQ 即常见问题库，它一般作为自动问答系统的子部分存在。比如“百度知道”，每当用户输入一个问题时，首先可以查找与之相似的问题及其对应的答案。……但是由于汉语表现形式的多样性，同样一个问题往往有多种表现形式，因此在 FAQ 中很难查找到一模一样的问句。

2.1 基于向量空间模型的句子相似度算法

向量空间模型（Vector Space Model，VSM）最初用在信息检索（IR）中用来对用户查询和语料库文档建模，如今已经得到了广泛的应用，如在句子或文档的表示中，就是通过句子中去掉停用词后剩下的有效词来构成向量空间，然后在该向量空间中将待计算的句子进行向量化，以两个向量夹角的余弦值作为句子之间的相似度度量。

2.2 实验方法

本文对相似度计算的结果评测方法选择 Pooling 方法评测， P@N 代表参与评测的算法都要返回前 N 个答案。分别采用准确率(Precision)、召回率(Recall)、F 值以及 MRR(Mean Reciprocal Rank)、MAP(Mean Average Precision)五个指标进行评价，其中 MRR、MAP 的公式如下：

$$PCD(P_k) = \frac{1}{m} \times \sum_{i=1}^{m} \frac{f_i(p_{k+1}) - f_i(p_{k-1})}{f_i^{\max} - f_i^{\min}} \quad (1)$$

$$F(x) = \partial \times n(x) + PCD(x) \quad (2)$$

其中，$f_i(p_{k+1})$、$f(p_k)$、$f_i(p_{k-1})$为连续的3个解的各个目标函数值；$PCD(x)$为解 x 的拥挤距离，∂为支配数目相对于拥挤距离的权值。

2.3 实验结果

本文选择的测试问句如表1所示，并分别返回前5、10、20个答案作为该算法返回的结果。

表1 测试问句

问句号	问句
S-1	皮肤偏黑，什么颜色的更适合？
S-2	谁知道有什么生发的方法？
S-3	直肠癌化疗后掉头发怎么办？
S-4	脂肪粒形成的主要原因是什么？

批注 [A20]: 公式的处理：
1、使用 Mathtype 公式编辑器编辑，不以图片方式在文档中插入公式；
2、按（1）、（2）全文编号；
3、在文中引用时用：式(1)、式(2)…

批注 [A21]: 表随正文，先见文字后见表，标注：“如表 1 所示”或“（见表 1）”。

批注 [A22]: 表的处理：
1、按表 1、表 2……全文编号；
2、表名齐备；
3、使用能够直接编辑的三线表。

S-5　　脸部皮肤不细腻，用什么洗面奶？

如图1所示，语义与统计相融合的词语相似度算法(M-3)在不考虑检索到结果的相关度排序问题下，效果最好，其次为基于语义的词语相似度算法(M-2)，最后是基于统计的词语相似度算法(M-1)。

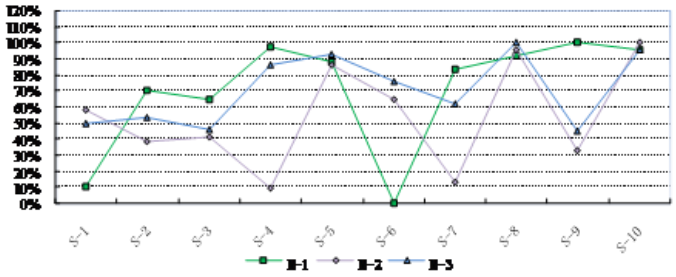

图 1　P@10 的 MAP 对比

3　结语

可以看出，基于统计的词语相似度受制于训练语料的规模，由于数据稀疏（止鼾器出现次数很少），由统计方法计算“止鼾器”的结果不理想，而基于语义的词语相似度算法可以得到较好的效果；从第三个问题（S-3）“直肠癌化疗后掉头发怎么办？”可以看出，单纯的基于语义计算词语相似度则完全依赖语义资源，语义词典中“直肠癌”与“头发”的相似度较低，而基于统计的词语相似度算法则能给出较高的相似度值。可见，将两者融合能有效地克服各自算法的缺点，给出更加合理的词语相似度数值。

参考文献：

期刊

作者. 论文题目[J]. 刊名，年，卷（期）：起始页码-终止页码.

[1]杨寸月，郑鲁腾.蓝牙技术的优势与前景[J].软件导刊，2012，11(6)：45-46.

[2]TURNEY P D. Similarity of semantic relations [J]. Computational Linguistics Journal, 2010, 32(3): 379-416.

[3] HEWITT J A. Technical services in 1983[J]. Library Resource Services, 1984, 28(3):205-218.

专著

作者. 书名[M]. 出版地：出版社，出版年.

[4]苗夺谦，李德毅，姚一豫.不确定性与粒计算[M].北京：科学出版社，2011.

[5] CRAWFORD W, GORMAN M. Future libraries: dreams, madness, & reality[M]. Chicago: American Library Association, 1995.

译著

作者. 书名[M]. 译者,译. 出版地：出版社，出版年.

[6]刘兵.Web 数据挖掘[M].余勇，译.北京：清华大学出版社，2012.

批注 [A23]: 图随正文，先见文字后见图，文中标注：“如图 1 所示”或“（见图 1）”

批注 [A24]: 图的处理：
1、按图 1、图 2……全文编号；
2、图名齐备；
3、尽量提供可供编辑的原图；
4、导出的图在字体、字号、分辨率、格式等方面作要求（中文字体：宋体，普通黑色；英文字体：Times New Roman；字号：7pt；图片宽度不超过 86mm，分辨率 300 像素以上；图片格式：TIF；无需底色或背景色）。

批注 [A25]: 参考文献：
1、据实使用，一般**不少于 20 条**；
2、尽可能**查全、查新**，涵盖**国内文献**和**国外文献**；
3、在正文中按出现的先后顺序**实引**（用黄色标记）；
4、被多次引用的文献，按首次出现的顺序编码，在文后只列一次。

批注 [A26]: 期刊类文献按“年，卷（期）：起止页码.”写全；
特殊情况下（部分英文期刊文献）可只写“年（期）：起止页码.”或“年，卷：起止页码.”

批注 [A27]: 按[1][2]…标号，与正文中的实引序号对应

批注 [A28]: 英文文献中的人名：
1、全部大写（只取前 3 个即可，余下用“，et al.”）；
2、姓在前、名在后，姓全拼、名用首字母缩写；
3、省略所有缩写点。
如引用 Rinku Dewri 的文章时，参考文献中的格式为 DEWRI　R; 当出现第二、三作者时，格式同第一作者。

批注 [A29]: 英文题名：
第一个单词首字母大写，其余为小写（专有名词等除外）。

批注 [A30]: 英文文献/会议名：
每一个单词首字母大写。

会议论文

作者. 论文题目[C].地点：会议名，年份. 或

作者. 论文题目[C].文集/会议名，年份：起始页码-终止页码.

[7] SPIVAK G C. Can the subaltern speak[C]. Victory in Limbo: Imigism, 2010:271-313.

学位论文

作者. 论文题目[D]. 所在城市：保存单位，年份.

[8]应晓敏. 面向 Internet 个性化服务的用户建模技术 [D] . 长沙：国防科技大学，2003.

技术标准

起草责任者，技术标准代号顺序号—发布年. 技术标准名称[S]. 出版地：出版社，出版年. 或

技术标准代号顺序号—发布年. 技术标准名称[S].

[9]国家标准局信息分类编码研究所. **GB / T2659-1986.世界各国和地区名称代码[S].**北京:中国标准出版社，1988.

专利

申请者. 专利名[P]. 国名,专利号，发布日期.

[10]刘加林. 多功能一次性压舌板[P].中国，92214985. 2,1993-04-14.

科技报告

作者.文题[R].报告题名，报告代码及编号，年份. 或

作者.文题[R].地名：责任单位，报告代码及编号，年份.

[11]BOZEMAN B. Knowledge value collectives: the proof of science in the putting [R].Contractor Report, AIAA-98-4484，2012.

报纸文章

作者. 文题[N]. 报纸名，出版日期（版次）.

[12]张田勤. 建设网络强国须关注企业信息安全[N]. 电脑报，2014-03-26(2).

电子文献

作者. 文题[EB/OL]. http://……

[13] 刘群，李素建.基于知网的词汇相似度计算[EB/OL].http://www.keenage.com.

数据库文献

作者. 文题[DB/OL].数据库名：编码，年份.

[14] XIA Y, QIN T, CHEN W, et al. Dual supervised learning [EB/OL]. Arxiv Preprint:1707.00415，2017.

网络优先出版文献

作者. 文题[J/OL].刊名.网络出版时间.网络出版地址.

[15] 李亚欣,蔡永香,邓舒颖.基于 Swift 对 Objective-C 开发的移动应用程序优化[J/OL].软件导刊：39-43.2018-07-17. http://kns.cnki.net/kcms/detail/42.1671.TP.20180717.1334.130.html.

批注 [A31]: 此类为**优先数字出版文献**。**《软件导刊》现已实施优先数字出版**，提前纸刊 3 个月左右在中国知网进行**单篇优先数字出版**。

其中，网络出版时间和网络出版地址在文献首页有显示。欢迎引用！

（注意：投稿请认准软件导刊官方网站 http://www.rjdk.org，谨防假冒）