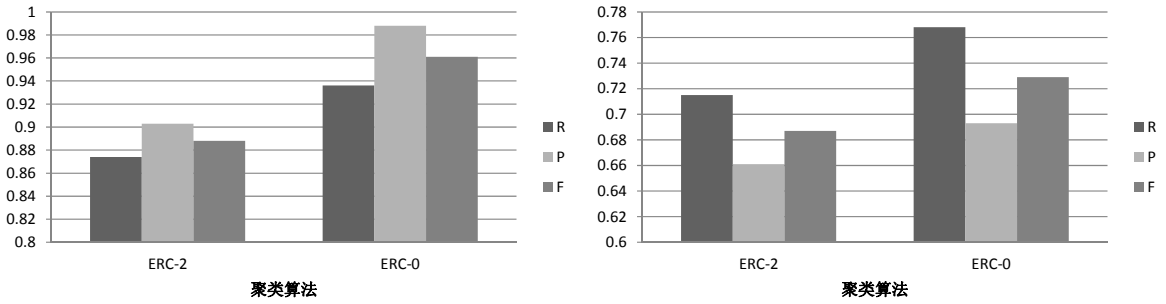


(a) Cora (b) A-G

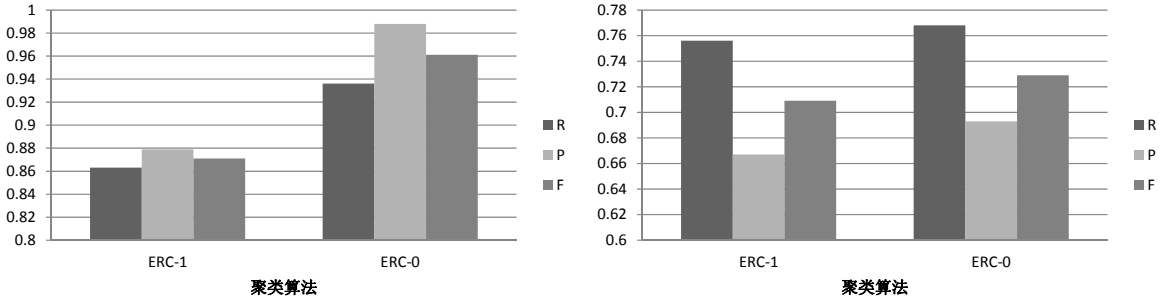
Fig.7 Comparisons of the two proposed cluster-vertex similarities on real datasets

图 7 两种簇点相似度在真实集上的对比

4.2.4 数据对象排序测试

本节测试数据对象排序在 ERC 算法中的作用.首先在两个真实数据集上进行对比实验,然后在不同分布类型的生成数据集上进行对比实验.

在两个真实数据集上,测试数据对象排序在 ERC 算法中的作用,将没有数据对象排序的 ERC 算法记作 ERC-1.图 8(a)中:在 Cora 数据集上,ERC-0 的 F 值比 ERC-1 高 10.3%;图 8(b)中:在 A-G 数据集上,ERC-0 的 F 值比 ERC-1 高 2.8%.在两个真实数据集上,ERC-0 的聚类效果都要比 ERC-1 好.可见,ERC 算法中数据对象排序有利于提高聚类结果的精确性.分析原因,两个真实数据集都是非均匀分布的,根据定理 3,将数据对象按照信用度降序排列的聚类结果比随机排列的更好.另外,比较不同数据集上的ΔF发现,在 Cora 数据集上的ΔF要比 A-G 数据集上的ΔF要大很多.从两者的数据分布来看:Cora 中真实类簇的大小差别很大,A-G 中真实类簇的规模以 2~3 为主.越不均匀分布的数据集上,数据对象排序带来的聚类顺序的变化越大,对聚类结果的影响也越大.后面将在生成数据集上专门针对数据分布的不均匀程度进行实验.



(a) Cora (b) A-G

Fig.8 Tests of data objects ordering's influence on clustering on real datasets

图 8 在真实集上测试数据对象排序对聚类的影响

在不同数据分布的生成数据集上,测试数据对象排序在聚类中的作用.用 UIS 数据生成器分别生成均匀分布的数据集、zipf 分布的数据集和泊松分布(λ=3)的数据集,规模都是 3 000 条.均匀分布记作 unf,泊松分布记作 psn.图 9 中:在均匀分布的数据集上,ERC-0 的 F 值与 ERC-1 几乎相当;在 zipf 分布的数据集上,ERC-0 的 F 值比 ERC-1 高 22.9%;在泊松分布的数据集上,ERC-0 的 F 值比 ERC-1 高 5.4%.首先,均匀分布的数据集上,ERC-1 和 ERC-0 的表现几乎相同,数据对象排序在均匀分布的数据上对聚类结果的影响几乎没有;其次,在 zipf 分布的和泊松分布的数据集上,ERC-0 的效果明显好于 ERC-1,数据对象排序在非均匀的数据分布上能够提高聚类结果的精确性;最后,在 zipf 分布数据集上的ΔF要远大于在泊松分布数据集上的ΔF.分析原因,zipf 分布的数据集上,真实类簇大小的不均匀程度远大于泊松分布数据集上的不均匀程度.

为了进一步测试数据分布的不均匀程度对数据对象排序在 ERC 算法中的影响,生成一组泊松分布的数据集,调节λ来控制分布变化,λ越大,分布悬殊越大,λ分别取 1,2,3,4 和 5,数据规模都是 3 000 条.图 10 中:整体来讲,ERC-0 的 F 值高于 ERC-1;具体来讲,随着泊松分布的参数λ增长(从 1~5),ERC-0 和 ERC-1 的ΔF 在不断增大,F 值提高率从 0.9%增长至 12.2%.综上所述,在越不均匀分布的数据集上,数据对象排序对 ERC 算法的聚类结果的精确性提高越大.

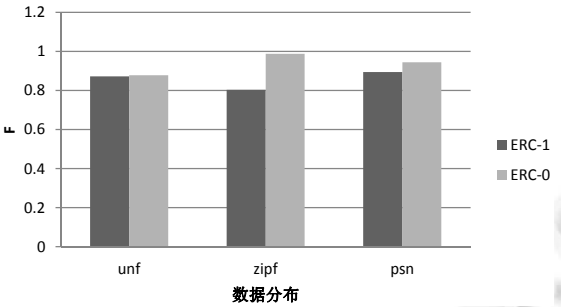


Fig.9 Data objects ordering's influences on clustering with different data distributions

图 9 不同数据分布下数据对象排序对聚类的影响

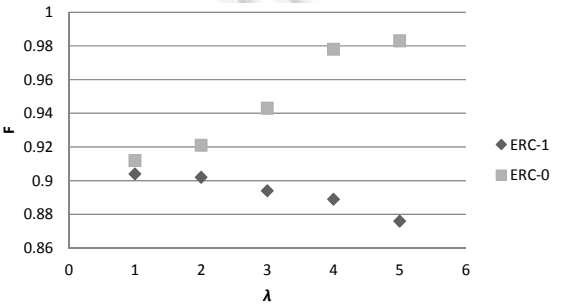


Fig.10 Data objects ordering's influence on clustering with Poisson distributions of different biases

图 10 数据对象排序在不均匀程度不同的泊松分布下对聚类结果的影响

4.2.5 计算优化测试

本节测试计算优化对 ERC 算法的作用.表 1 中,ERC-3 算法与 ERC-0 算法的不同之处在于,ERC-3 算法没有计算优化,聚类过程中需要迭代地计算簇点相似度.首先比较两个算法的计算结果的精确性,ERC-0 算法与 ERC-3 算法的计算结果是相同的,图 11 是两种算法在 3 个真实数据集上的结果对比.计算优化对 ERC 算法计算结果的精确性没有影响.

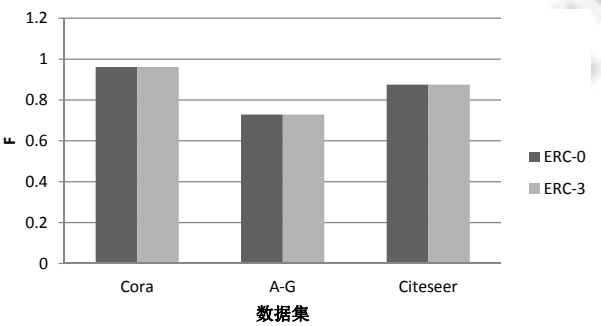


Fig.11 Tests of computation optimization's influence on accuracy with real datasets

图 11 在真实数据集上测试计算优化对精确性的影响

然后比较两者的时间开销,分别在 3 个真实的数据集上进行实验,得到结果见表 3.根据表 3,ERC-0 算法比 ERC-3 算法在 Cora,A-G 和 Citeseer 这 3 个数据集上分别少开销 38.5%,37.4%和 33.3%的时间,因此,计算优化可以为 ERC 算法带来较大的时间开销节省.基本的单例簇点相似度在对记录进行排序的时候已经计算得到;利用计算优化,ERC-0 算法在聚类过程只需要对基本的单例簇点相似度进行线性组合,即可计算出类簇与结点的相似度,因而节省了较大的开销.

Table 3 Tests of computation optimization's influence on time cost with real datasets (s)

表 3 在真实数据集上测试计算优化对时间开销的影响 (s)

	Cora	A-G	Citeseer
ERC-0	0.837	0.615	68.63
ERC-3	1.362	0.983	102.95

5 相关工作

实体识别是数据质量的一个重要方面,又称为实体解析、实体匹配、记录匹配、实体辨析、合并与清洗等[2−14].实体识别主要包括对象相似度计算和匹配决定两个必要组成部分和一个可选组成部分:分块.已有工作针对对象相似度计算提出很多相似度算法,以适应不同类型的数据对象[8,9,12−14],如文本的相似度算法、基于实体关系的相似度算法等.大数据实体识别中,分块技术变得不可或缺.分块技术通过减小搜索空间来降低开销[16−20],提高大数据实体识别效率.对象相似度计算和分块技术在引言和第 1.1 节中描述,此处不再赘述.

实体识别中的匹配决定方法按照是否需要训练过程可分为监督类方法和非监督类方法:监督类方法要求用户提供高质量的标注数据来训练分类器,然后利用分类算法来执行匹配决定[2,3,5](见第 1.1 节);监督类方法严重依赖领域知识,造成其应用的局限性.传统的非监督类方法通过测定相似度阈值来判定候选对是否匹配.Hassanzadeh 等人将已有的聚类算法应用在匹配决定中,并通过对比实验证明,基于聚类的方法比基于阈值的方法更有效[4].

Center 算法[24]首先将相似对降序排列依次加入队列;遍历队列,从第 1 个扫描到的候选对中选一个结点 v_i 作为下一个类簇的中心;将后续所有与 v_i 相似的结点加入到最新生成的类簇,并将包含这些结点的候选对从队列中删除;不断循环,直到生成聚类结果.即,所有的类簇都由一个中心和与它相近的结点组成.Center 算法在网络文档检索上非常高效.文献[4]在 Center 算法基础上提出 Merge-Center 算法,它允许足够相似的类簇合并.在实体识别的匹配决定中,Merge-Center 算法比 Center 算法更高效[4].Markov Clustering 算法[25]通过模拟图上的马尔可夫随机流来进行聚类,它的基本思想是:图上一块关联紧密的区域会形成一个类簇,类簇内部的流的总量会很强;相反,两个类簇之间的的关联较弱,类簇间的流的总量会较弱.Markov Clustering 算法在图上进行随机游走,加强原本就强的流(即簇内),减弱原本就弱的流(即簇间),不断迭代,直到类簇结构形成.Markov Clustering 算法被应用在生物信息领域中,并能快速地得到高质量的聚类结果[26].MinCut 聚类算法[27]在图上发现边的最小切割来实现聚类,该算法基本思想是:发现类簇间的最小切割,从而使得簇内的边的权重和最大化,这样得到的聚类结果将是簇内紧密耦合,簇间松散关联.MinCut 算法被应用在引文数据和网络数据聚类中[28].Articulation Point Clustering 算法在图上发现关节点和重连通分量来进行聚类[29],每个重连通分量就是一个类簇.该算法被应用于发现博客圈中热点话题[32].Hassanzadeh 等人首次将上述聚类算法应用于匹配决定中,并进行了实验对比[4].通过本文的实验对比可知:在 Cora 和 A-G 两个真实的数据集上,本文提出的 ERC 算法的识别效果,要比 Merge-Center,Markov Clustering,MinCut 和 Articulation Point Clustering 等算法的识别效果更好.

大数据的一个特点是产生和更新速度快,Gruenheid 等人在已有算法的基础上提出了增量的实体识别算法,能增量、快速地处理逐步更新的数据[33].为满足实时的应用需要,即,在短时间内识别大部分的数据对象,提出了 Pay-as-you-go 实体识别算法[6,7].基于时间特征的实体识别算法通过分析数据对象随时间的演化信息来完成识别任务[8,9].

随着众包(crowdsourcing)的流行,一些研究工作借助大众的力量提出了基于众包的实体识别算法[10,11].

6 结束语

实体识别对于数据集成和数据挖掘都必不可少.本文针对非监督的实体识别中匹配决定问题,提出了一个基于随机游走模型的聚类算法——ERC 算法.该算法利用图上的随机游走,通过挖掘图结构来计算聚类过程中类簇和结点的相似度;为了优化聚类顺序、提高识别结果的精确性,该算法根据数据对象的信用度来进行降序排列.通过在两个真实数据集上和若干生成数据集上的实验对比和分析,验证了 ERC 算法的有效性以及其组成

部分的作用.在未来的工作中,作者将致力于提出更加理论化的阈值确定方法.另外,如何使本文提出的算法能够更高效、准确地处理增量实体识别中匹配决定,也将是进一步的研究工作.

References:

[1] Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston: Houghton Mifflin Harcourt, 2013.

[2] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans. on Knowledge and Data Engineering, 2007,19(1):1−16. [doi: 10.1109/TKDE.2007.250581]

[3] Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems. Proc. of the VLDB Endowment, 2010,3(1-2):484−493. [doi: 10.14778/1920841.1920904]

[4] Hassanzadeh O, Chiang F, Lee HC, Miller RJ. Framework for evaluating clustering algorithms in duplicate detection. Proc. of the VLDB Endowment, 2009,2(1):1282−1293. [doi: 10.14778/1687627.1687771]

[5] Guo ZM, Zhou AY. Data quality and data cleaning: A survey. Ruan Jian Xue Bao/Journal of Software, 2002,13(11):2076−2028 (in Chinese with English abstract).

[6] Altowim Y, Kalashnikov DV, Mehrotra S. Progressive approach to relational entity resolution. Proc. of the VLDB Endowment, 2014,7(11):999−1010. [doi: 10.14778/2732967.2732975]

[7] Papenbrock T, Heise A, Naumann F. Progressive duplicate detection. IEEE Trans. on Knowledge and Data Engineering, 2015,27(5): 1316−1329. [doi: 10.1109/TKDE.2014.2359666]

[8] Chiang YH, Doan AH, Naughton JF. Tracking entities in the dynamic world: A fast algorithm for matching temporal records. Proc. of the VLDB Endowment, 2014,7(6):469−480. [doi: 10.14778/2732279.2732284]

[9] Li F, Lee ML, Hsu W, Tan WC. Linking temporal records for profiling entities. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2015. 593−605. [doi: 10.1145/2723372.2737789]

[10] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. Proc. of the VLDB Endowment, 2014,7(12): 1071−1082. [doi: 10.14778/2732977.2732982]

[11] Wang S, Xiao X, Lee CH. Crowd-Based deduplication: An adaptive approach. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2015. 1263−1277. [doi: 10.1145/2723372.2723739]

[12] Benjelloun O, Garcia-Molina H, Menestrina D, Su Q, Whang SE, Widom J. Swoosh: A generic approach to entity resolution. The Int'l Journal on Very Large Data Bases, 2009,18(1):255−276. [doi: 10.1007/s00778-008-0098-x]

[13] Bhattacharya I, Getoor L. Collective entity resolution in relational data. ACM Trans. on Knowledge Discovery from Data, 2007, 1(1):5. [doi: 10.1145/1217299.1217304]

[14] Sun CC, Shen DR, Kou Y, Nie TZ, Yu G. A related data oriented joint entity resolution approach. Chinese Journal of Computers, 2015,38(9):1739−1754 (in Chinese with English abstract).

[15] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. In: Getoor L, Senator TE, Domingos PM, Faloutsos C, eds. Proc. of the ACM KDD Workshop on Data Cleaning and Object Consolidation. New York: ACM Press, 2003. 73−78.

[16] Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. on Knowledge and Data Engineering, 2012,24(9):1537−1555. [doi: 10.1109/TKDE.2011.127]

[17] Papadakis G, Papastefanatos G, Koutrika G. Supervised meta-blocking. Proc. of the VLDB Endowment, 2014,7(14):1929−1940. [doi: 10.14778/2733085.2733098]

[18] Fisher J, Christen P, Wang Q, Wang Q, Rahm E. A clustering-based framework to control block sizes for entity resolution. In: Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2015. 279−288. [doi: 10.1145/2783258.2783396]

[19] Karakasidis A, Koloniari G, Verykios VS. Scalable blocking for privacy preserving record linkage. In: Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2015. 527−536. [doi: 10.1145/2783258.2783290]

[20] Kenig B, Gal A. Efficient entity resolution with mfiblocks. Proc. of the VLDB Endowment, 2009,4(1-2):484−493.

[21] Arasu A, Götz M, Kaushik R. On active learning of record matching packages. In: Elmagarmid AK, Agrawal D, eds. Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2010. 783−794. [doi: 10.1145/1807167.1807252]

[22] Xu R, Wunsch I. Survey of clustering algorithms. IEEE Trans. on Neural Networks, 2005,16(3):645−678. [doi: 10.1109/TNN.2005.845141]

[23] Sun JG, Liu J, Zhao LY. Clustering algorithms research. Ruan Jian Xue Bao/Journal of Software, 2008,19(1):48−61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm

[24] Haveliwala T, Gionis A, Indyk P. Scalable techniques for clustering the Web. In: Proc. of the 2000 Int'l Workshop on the Web and Databases. New York: ACM Press, 2000. 129−134.

[25] Dongen S. Graph clustering by flow simulation [Ph.D. Thesis]. Utrecht: University of Utrecht, 2000.

[26] Brohee S, Van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics, 2006, 7(1):488. [doi: 10.1186/1471-2105-7-488]

[27] Flake GW, Tarjan RE, Tsioutsiouliklis K. Graph clustering and minimum cut trees. Internet Mathematics, 2004,1(4):385−408. [doi: 10.1080/15427951.2004.10129093]

[28] Cormen TH, Leiserson CE, Rivest RL. Introduction to Algorithms. Cambridge: MIT Press, 1990.

[29] Bansal N, Chiang F, Koudas N, Tompa FW. Seeking stable clusters in the blogosphere. Proc. of the VLDB Endowment, 2007: 806−817.

[30] Motwani R, Raghavan P. Randomized Algorithms. Cambridge: Cambridge University Press, 1995.

[31] Tong H, Faloutsos C, Pan JY. Fast random walk with restart and its applications. In: Proc. of the 6th IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2006. 613−622. [doi: 10.1109/ICDM.2006.70]

[32] Clauset A, Shalizi CR, Newman ME. Powerlaw distributions in empirical data. SIAM Review, 2009,51(4):661−703. [doi: 10.1137/070710111]

[33] Gruenheid A, Dong XL, Srivastava D. Incremental record linkage. Proc. of the VLDB Endowment, 2014,7(9):697−708. [doi: 10.14778/2732939.2732943]

附中文参考文献:

[5] 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报,2002,13(11):2076−2028.

[14] 孙琛琛,申德荣,寇月,聂铁铮,于戈.面向关联数据的联合式实体识别方法.计算机学报,2015,38(9):1739−1754.

[23] 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48−61. http://www.jos.org.cn/1000-9825/19/48.htm

孙琛琛(1987－),男,山西平遥人,博士生,CCF 学生会员,主要研究领域为实体识别.

聂铁铮(1980－),男,博士,副教授,CCF 会员,主要研究领域为数据质量,数据集成.

申德荣(1964－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为分布式数据管理,数据集成.

于戈(1962－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据库,大数据管理.

寇月(1980－),女,博士,副教授,CCF 会员,主要研究领域为实体搜索,数据挖掘.