北京大学学报(自然科学版) 第52卷 第1期 2016年1月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 1 (Jan. 2016)
doi: 10.13209/j.0479-8023.2016.003

基于主题敏感的重启随机游走实体链接方法

李茂林

北京邮电大学智能科学与技术中心, 北京 100876; E-mail: mlli@bupt.edu.cn

摘要 实体链接任务的目的是将文本中的实体指称链接到知识库中与之对应的无歧义实体。针对此任务, 提出一种基于主题敏感的重启随机游走的实体链接方法。该方法首先使用实体指称的背景文本信息将实体指称扩充为全称, 并在维基百科知识库中搜索候选实体, 得到候选实体集合; 根据上述中间结果构建图, 利用在图上的主题敏感重启随机游走得到的平稳分布对候选实体集合进行排序, 选出 top 1 的候选实体作为目标实体。实验结果表明, 该方法在 KBP2014 实体链接数据集上实验的 F 值为 0.623, 高于其他系统实验的 F 值, 能够有效提高实体链接系统的整体性能。

关键词 实体链接; 随机游走; 维基百科

中图分类号 TP391

An Entity Linking Approach Based on Topic-Sensitive Random Walk with Restart

LI Maolin

Center for Intelligence Science and Technology, Beijing University of Posts and Telecommunications, Beijing 100876; E-mail: mlli@bupt.edu.cn

Abstract Entity linking is the process of linking name mentions in text with their referent entities in a knowledge base. This paper tackles this task by proposing an approach based on topic-sensitive random walk with restart. Firstly, the context information of mentions is used to expand mentions and search the candidate entities in Wikipedia knowledge base for mentions. Secondly, graph can be constructed in accordance with the intermediate result in the pre step. Finally, the topic-sensitive random walk with restart model is used to rank the candidate entities and choose the top 1 as the linked entity. Experimental results show that proposed approach on KBP2014 data set gets F score 0.623 which is higher than every other systems' mentioned in this paper. The proposed approach can improve the entity linking system's performance.

Key words entity linking; random walk; Wikipedia

实体链接(entity linking)是将文本中的实体指称链接到知识库中一个无歧义实体的过程。随着信息技术的发展, 网络上产生了大量的非结构化文本数据, 使用实体链接技术有利于从这些大量的非结构化文本中挖掘有价值的信息, 对于计算机理解文本的真实含义有重要影响。此外, 实体链接技术也有利于共指消解、文本分类、用户兴趣发现以及推荐系统等方面的研究[1]。

通过分析 KBP2014 实体链接数据集以及维基百科知识库, 发现目前实体链接面临的两个主要问题。1)在网络中产生的大量非结构化文本中广泛存在实体多样性和歧义性现象[2]。实体多样性指一个实体可以用多个名称表示, 实体歧义性指一个名称可以代表多个实体。2)在实体链接的候选实体排序过程中, 前人没有考虑实体指称和所对应背景文本的主题倾向。例如, 当实体指称为"Apple"时, 没有考虑实体指称本身是更倾向于水果"Apple"还是科技领域的"Apple"公司及相关产品。

收稿日期: 2015-06-07; 修回日期: 2015-08-28; 网络出版日期: 2015-09-29

本文通过分析目前实体链接面临的主要问题以及前人工作的缺陷，提出下列方法：1)利用 LDA (latent dirichlet allocation)主题模型，生成一个实体名称多样性词表，利用此词表更好地适应实体歧义性现象，通过将其整合到候选实体生成模块，提高实体链接系统的整体性能；2)在前人工作的基础上，对图的构建方法进行改进，使其适应实体的多样性现象，并提出一种基于主题敏感的重启随机游走实体链接方法，使最终得到的候选实体集合的排序结果更加准确。

1 相关工作

1.1 实体链接

目前，已提出多种解决实体链接的方法，主要分为单一实体链接和协同实体链接两种类型。

单一实体链接方法在进行实体链接时，仅考虑当前正在处理的实体指称与候选实体之间是否存在对应关系。Mihalcea 等[3]提出一种基于词袋(bag of words)模型的实体链接方法，将实体指称的背景文本以及候选实体对应的文本转换为词袋向量，然后计算两者的余弦相似度，最终选择相似度最高的实体作为目标实体。Cucerzan[4]使用维基百科中的分类目录信息对词袋模型进行增强。Dredze 等[5]和 Zheng 等[6]通过学习排序算法进行实体链接，考虑了候选实体之间的相对位置。Pink 等[7]通过对与实体相关的局部文本信息进行建模，并使用有监督机器学习方法对命名实体进行链接。Dalton 等[8]提出一种邻近相关性推理数学模型进行实体链接。

单一实体链接方法在实体链接过程中每次只考虑当前正在处理的实体指称本身，并没有考虑与背景文本中其他实体指称之间的语义关系。协同实体链接方法通过建立全局语义的约束，考虑了当前处理的文本中所有实体指称之间的语义关系。Alhelbawy 等[9]利用背景文本中所有实体指称的候选实体构造图，并使用 PageRank 算法计算每个候选实体的权重，最终选择权重最高的实体作为最佳目标实体，但在构造图时只简单地根据节点的出度将转移概率平均分配。Han 等[10]对节点之间的转移概率计算方法进行优化，并构造了一种指示图，通过基于指示图的集体推理数学模型推断出最佳的目标实体。但是，他们的方法在计算候选实体之间的语义相关度时会出现负值，所以得出的语义相关度并不准确，并且构造的指示图会出现不为强连通图的情况，不能保证最终得到的平稳分布是合理的，会影响最终候选实体排序的准确性。

1.2 随机游走

随机游走(random walk)是一种数学统计模型，最早由 Pearson[11]提出。随机游走由一连串的轨迹组成，每一步的运动都是随机的，这种随机过程可用马尔科夫链表示，从一个点移动到另一个点的转移概率与时间无关。重启随机游走(random walk with restart)模型由 Grady[12]提出，最早用于图像分割。重启随机游走是一种特殊类型的随机游走，当将要进行下一步移动时有两种选择：一种是以一定概率根据状态转移矩阵随机地选择下一个状态，另一种是以一定的概率选择任意点开始随机游走。

PageRank 算法是由 Page 等[13]提出的一种基于随机游走的链接分析方法，用来衡量网页的重要性。然而 PageRank 算法忽略了输入查询词的主题倾向信息。为了解决此问题，Haveliwala[14]提出主题敏感的 PageRank 算法，通过预定义几个主题类别，在用户进行查询时确定其主题倾向，根据此主题倾向给出更合理的网页重要性排序结果。

本文参考了主题敏感的 PageRank 算法思想，并将其运用到实体链接中。

2 基于主题敏感的重启随机游走实体链接

基于主题敏感的重启随机游走实体链接方法主要包括预处理、实体指称扩充、候选实体生成、候选实体排序和实体聚类5个部分，该方法流程如图1所示。

2.1 预处理

预处理包括命名实体识别、基于维基百科的资源获取、LDA 主题模型的建立、实体名称多样性词表的生成以及维基百科知识库文本聚类。

2.1.1 命名实体识别

本文重点关注文本中的人名(Person)、地名(Location)以及组织机构名(Organization) 3 种类型的命名实体。通过使用 Stanford NER 工具①对实体指称的背景文本以及维基百科中实体的背景文本进行命名实体识别，并将其中的每个命名实体预处理

① http://nlp.stanford.edu/software/CRF-NER.shtml

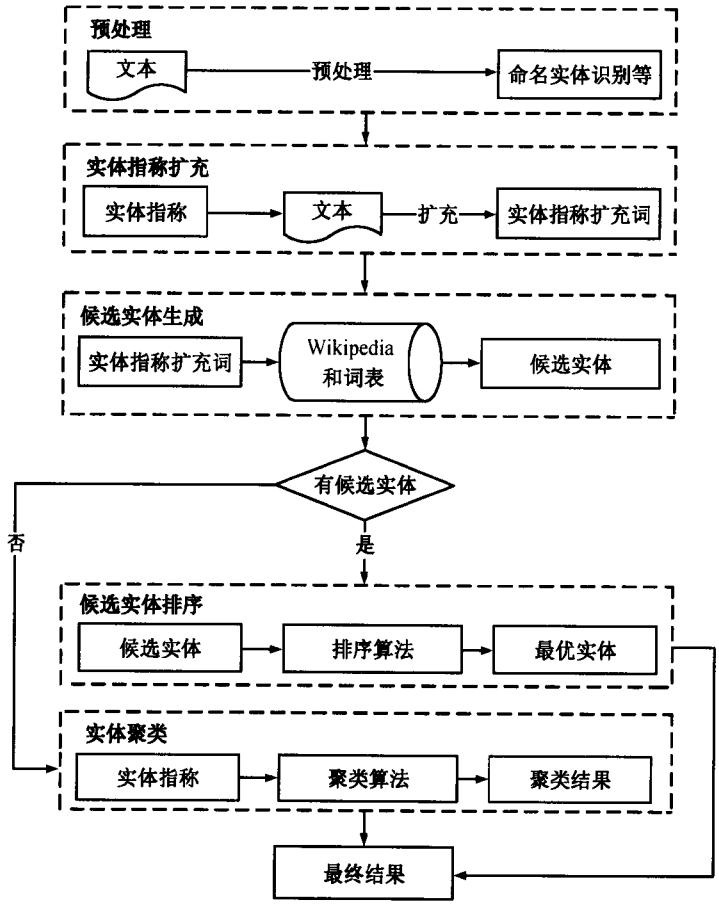

图 1　基于主题敏感的重启随机游走的实体链接架构

Fig. 1　Scheme of entity linking approach based on topic-sensitive random walk with restart

为一个单词，目的在于，一方面可以为此命名实体计算其在语料中的 TF-IDF 权重，另一方面为实体名称多样性词表的生成做准备。

2.1.2　基于维基百科的资源获取

在维基百科知识库中包含大量的实体及其对应的背景文本，从中可以挖掘出同一实体的简称、别名等其他表现形式，为候选实体生成步骤提供帮助。本文借鉴谭咏梅等[2]对维基百科的资源获取方法，从页面标题、重定向信息、锚文本以及消歧页面中获取资源，并对其进行补充和完善，通过解析维基百科知识库中的“fullname”与“nickname”字段得到实体的全名和昵称形式的实体指称。

2.1.3　LDA 主题模型的建立

LDA 主题模型用于实体名称多样性词表的生成以及维基百科知识库文本聚类。在命名实体识别预处理的基础上，去除停用词和低频词，针对维基百科知识库中的所有实体的背景文本，通过 gensim 工具①计算文本中每一个单词的 TF-IDF 权重，并建立 LDA 主题模型。

2.1.4　实体名称多样性词表的生成

实体名称多样性词表中存放了每个实体可能对应的多种实体指称。将实体名称多样性词表与基于维基百科获取的资源相结合，有利于解决实体链接中的实体歧义性问题，以提高实体链接系统的整体性能。

Bradford[15]提出一种基于 LSA (latent semantic analysis)的实体名称多样性词典构建方法，通过使用 LSA 模型中的单词向量之间的相似度计算和比较来提取每个实体不同的实体指称表示形式。本文首先通过命名实体识别，将命名实体预处理为一个单词，然后将 LSA 模型改用 LDA 主题模型生成单词向量，并对 Bradford 工作中的参数进行调整，最

① http://radimrehurek.com/gensim/

终生成实体名称多样性词表。

2.1.5 维基百科知识库文本聚类

维基百科知识库文本聚类用于为维基百科知识库中的每一个实体对应的背景文本分配簇标签，表示此文本的主题倾向。由于主题敏感的重启随机游走模型需要预先对维基百科知识库中的所有实体对应的背景文本进行文本分类，若此分类任务由人工完成会耗费大量的时间，因此本文通过使用LDA主题模型，将实体对应的背景文本转换为文档向量，然后使用 Mini-Batch K-Means 聚类算法[16]，对所有文档向量进行聚类，最终为每一篇文本分配簇标签。

由于维基百科知识库中实体数量庞大，与 K-Means 算法相比，Mini-Batch K-Means 算法的聚类效率更高，因此选择此聚类算法。

2.2 实体指称扩充

背景文本中的实体指称往往具有很大的歧义性，例如在文本中出现“J.D.”和“J. D. Collins”两个实体指称，两者表示同一个实体，若将“J.D.”扩充为“J. D. Collins”，则会减少实体指称的“J.D.”歧义性，可以取得更好的实体链接效果。

对实体指称进行扩充会减少实体指称的歧义性，主要针对首字母缩写词、简写形式的实体指称进行扩充。1)首字母缩写词：如果一个实体指称中的所有字母均为大写，则将其作为首字母缩写词，通过在文本中搜索是否含有首字母大写匹配的字符串来进行扩充。2)简写形式：若文本中存在“long(short)”或“short(long)”形式的字符串，则通过判断括号内外字符串的长度大小来对实体指称进行扩充。3)其他：若上述过程都无法对当前实体指称进行扩充，则通过搜索文本的命名实体识别结果来进行扩充。例如，若文本中出现实体指称“Lichnowy”，且命名实体识别结果中存在类型为 Organization 的命名实体“Gmina Lichnowy”，则将“Lichnowy”扩充为“Gmina Lichnowy”。

2.3 候选实体生成

候选实体生成是为背景文本中的每一个实体指称在知识库中找到其可能代表的候选实体集合。此步骤会将实体指称的扩充形式与基于维基百科获取的资源进行字符串匹配。若实体指称扩充形式与资源中的某一指称形式完全匹配，则将此指称形式对应的实体作为候选实体。如果无法在维基百科知识库的资源中找到候选实体，则在实体名称多样性词表中选择与此实体指称相似度最高的指称作为其变体名称，再次在资源中查找候选实体。

由于维基百科知识库中收录的实体数量有限，可能无法为所有的实体指称找到候选实体，此时将满足此条件的实体指称定义为无指代实体指称，以 NIL 表示，并将其加入 NIL 集合中，以便在后续的步骤中对其中所有的无指代实体指称进行聚类，将表示同一种实体的实体指称聚为一簇。

2.4 候选实体排序

候选实体排序包括实体指称与候选实体之间局部相关度的计算、候选实体之间语义相关度的计算以及主题敏感的重启随机游走实体排序3个关键步骤。本文参考 Han 等[10]的工作，针对其方法中的不足进行改进，提出一种基于主题敏感的重启随机游走实体链接方法，以提升实体链接系统的性能。

主题敏感的重启随机游走实体链接主要利用实体指称与候选实体之间的局部相关度以及候选实体之间的语义相关度构建图并计算状态转移概率矩阵，通过使用实体指称以及候选实体的主题倾向(即所属的簇标签)设置随机游走中的重启节点，计算最终的平稳分布，最终根据每一项的值对候选实体进行排序，选取 top 1 作为最佳的目标实体。

本文所用符号的说明：m 表示一个实体指称，e 表示维基百科知识库中的一个无歧义实体，c 表示实体指称或候选实体对应的簇中心，C 表示由所有簇中心组成的集合，$E(m)$表示实体指称 m 的候选实体集合，粗体小写字母表示向量，$G=(V, E)$表示以节点的集合 V 以及边的集合 E 构成的图，$\boldsymbol{T}$ 表示转移概率矩阵，$\boldsymbol{s}$ 表示初始分布向量，$\boldsymbol{r}$ 表示随机游走过程中的分布向量。

2.4.1 实体指称与候选实体之间局部相关度

实体指称与候选实体之间的局部相关度是每一个实体指称 m 对应的背景文本与候选实体 e 对应的背景文本之间的相似度。以一定窗口大小取实体指称 m 的周围单词作为背景文本，计算每一个单词的 TF-IDF 权重，并将其转换为词袋模型向量 $\boldsymbol{m}$。同理，将候选实体 e 在维基百科知识库中背景文本转换为向量表示 $\boldsymbol{e}$，实体指称 m 与候选实体 e 之间的局部相关度(CP)计算方法如式(1)所示：

$$\mathrm{CP}(m,e)=\frac{\boldsymbol{m}\cdot\boldsymbol{e}}{|\boldsymbol{m}||\boldsymbol{e}|}。\tag{1}$$

2.4.2 候选实体之间语义相关度

局部相关性的计算只考虑每个实体指称 m 单独

与候选实体之间的相关度。为了弥补此缺陷，需要进行候选实体之间的语义相关度计算。对于候选实体集合中候选实体 e_1 与 e_2，若 e_1 与 e_2 之间的相关度越大，则其语义相关度的值越大。前人提出一些语义相关度计算方法，其中 Han 等[10]和 Milne 等[17]使用如下语义相关度计算公式(SR):

$$\mathrm{NGD}(e_1,e_2)=\frac{\log(\max(|S(e_1)|,|S(e_2)|))-\log(|S(e_1)\cap S(e_2)|)}{\log(|W|)-\log(\min(|S(e_1)|,|S(e_2)|))},\quad(2)$$

$$\mathrm{SR}(e_1,e_2)=1-\mathrm{NGD}(e_1,e_2),\quad(3)$$

其中，$S(e_1)$ 与 $S(e_2)$ 表示维基百科中所有页面中出现实体 e_1 与 e_2 的实体集合，W 表示维基百科中所有的实体。式(2)为标准化 Google 距离，表示两个实体之间的语义距离[18]。式(3)为 e_1 与 e_2 两个实体之间的语义相关度。

由式(2)和(3)可以得出，若 $S(e_1)$ 与 $S(e_2)$ 之间的交集越大，则 e_1 与 e_2 之间的语义相关度越大。由于式(2)的值域为 $(-\infty,1]$，当 e_1 与 e_2 的语义相关度为负值时，会影响图节点之间的转移概率的计算。这是因为转移概率不可能为负值，否则会造成候选实体排序结果不准确。因此，本文针对两个候选实体之间的语义相关度计算公式进行改进:

$$\mathrm{SR}(e_1,e_2)=\frac{1}{k\cdot\mathrm{NGD}(e_1,e_2)+1},\quad(4)$$

其中，k 为可调参数，取值为 0.1。式(4)通过取倒数的方式，将两个候选实体之间的语义距离转换为语义相关度，避免了语义相关度为负值的情况。

2.4.3　随机游走候选实体排序

随机游走候选实体排序主要包括图的构建以及主题敏感的重启随机游走候选实体排序两部分。

在前面工作的基础上，本文构建权重有向强连通图 $G=(V,E)$，其中 V 包含文本中所有的实体指称 m 和候选实体 e 作为节点，E 包含实体指称与候选实体的局部相关度以及候选实体之间的语义相关度信息，最终构建的图如图 2 所示。

图 2 中，m_1，m_2 和 m_3 表示实体指称，e_1，e_2 和 e_3 表示候选实体，图中的边为有向边，边的权重分别表示实体指称与候选实体之间的局部相关度以及候选实体之间的语义相关度，例如 $\mathrm{CP}(m_1,e_1)=0.6$，$\mathrm{SR}(e_1,e_2)=0.5$，其他边的权重依此类推。图中有灰色背景的节点表示其主题倾向属于同一簇。

图 2 中共有 3 种边，分别为实体指称 m 指向候选实体 e 的边、候选实体 e 之间互相指向的边以及候选实体 e 指向实体指称 m 的边。这 3 种边的转移概率计算公式分别为

$$P(m\to e)=\frac{\mathrm{CP}(m,e)}{\sum_{e\in N_m}\mathrm{CP}(m,e)},\quad(5)$$

$$P(e_i\to m)=\frac{\mathrm{CP}(m,e_i)}{\sum_{m\in N_{e_i}}\mathrm{CP}(m,e_i)+\sum_{e\in Ne_i}\mathrm{SR}(e_i,e)},\quad(6)$$

$$P(e_i\to e_j)=\frac{\mathrm{SR}(e_i,e_j)}{\sum_{m\in N_{e_i}}\mathrm{CP}(m,e_i)+\sum_{e\in Ne_i}\mathrm{SR}(e_i,e)}。\quad(7)$$

式(5)中，N_m 表示图中与实体指称 m 相邻的所有候选实体 e 的集合。由式(5)可得出，实体指称 m 指向候选实体 e 的转移概率适应了实体歧义性现象，如图 2 中，实体指称 m_1 指向候选实体 e_1 的转移概率为 0.46，指向候选实体 e_2 的转移概率为 0.54。

式(6)和(7)中，N_{e_i} 表示图中与候选实体 e_i 相邻的所有实体指称 m 与候选实体 e 的集合。由式(6)可以得出，候选实体 e_i 指向实体指称 m 的转移概率适应了实体多样性现象，例如图 2 中，候选实体 e_i 指向实体指称 m_2 和 m_3 的转移概率分别为 0.73 和 0.27。式(7)表示候选实体之间的转移概率。

由式(5)~(7)可以计算G上的转移概率矩阵 $\boldsymbol{T}$。

G 上的随机游走初始分布 $\boldsymbol{s}$ 为$|V|\times1$ 的向量，其

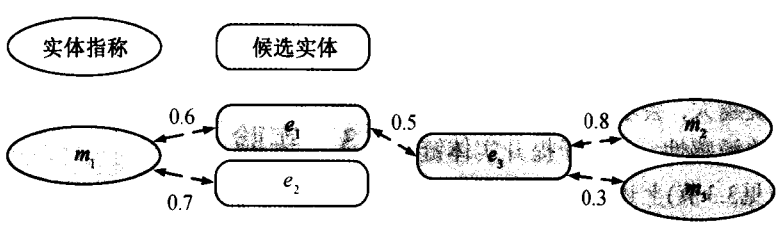

图 2　图结构

Fig. 2　Graph structure

值由两部分组成，分别为实体指称 m 与候选实体 e 的初始值，其中实体指称 m 的初始值的计算如式(8)所示：

$$s_{m_i}=\frac{\text{TF-IDF}(m_i)}{\sum_{m\in V_m}\text{TF-IDF}(m)}, \tag{8}$$

s_{m_i} 表示向量 s 中，实体指称 m_i 所对应的初始值，V_m 表示 G 中的所有实体指称 m 集合，TF-IDF(m_i)表示 m_i 的 TF-IDF 权值。可以看出，权值越大的实体指称 m 在随机游走过程中的影响越大。

为了确定图中每个实体指称 m 的主题倾向，将实体指称 m 对应的背景文本根据 LDA 主题模型转换为向量表示，然后使用 KNN (K-NearestNeighbor)算法分别计算其与每一个簇中心 c 之间的距离，选取距离最小的簇中心 c 所对应的簇标签作为实体指称 m 的簇标签，表示其主题倾向，如式(9)所示：

$$m.c=\arg\min_{c_i\in C}(\text{distance}(c,c_i)), \tag{9}$$

其中 $m.c$ 表示实体指称 m 主题倾向对应的簇中心，distance(c,c_i)表示计算两个簇中心 c 与 c_i 之间的欧氏距离。

为了使整个随机游走过程对实体指称 m 主题倾向敏感，当确定实体指称 m 的簇中心 $m.c$ 后，仅对 s 中与实体指称 m 具有相同簇中心即主题倾向的候选实体 e 的值进行初始化，其他候选实体的初始值设置为 0，如式(10)所示：

$$\begin{cases} s_{e_i}=\text{Indegree}(e_i)+\text{Outdegree}(e_i), \\ e_i\in\{e\mid e.c=m.c, e\in V\}, \\ s_{e_i}=0,\ e_i\in\{e\mid e.c\neq m.c, e\in V\}, \end{cases} \tag{10}$$

其中，s_{e_i} 表示向量 s 中候选实体 e_i 对应的初始值，$e.c$ 表示候选实体 e 对应的簇中心，Indgree(e_i)与 Outdegree(e_i)分别表示计算 e_i 的入度与出度。上述初始化方式基于两个假设：1)在重启随机游走过程中，在与实体指称 m 具有相同簇中心 $m.c$ 的候选实体 e 节点上发生重启，会使最终的平稳分布对主题更为敏感；2)一个候选实体 e 节点的度越大，在此节点发生重启的概率就越大，会增强重启随机游走最终平稳分布对主题的敏感性，有利于提升实体链接的性能。实验证明(见3.3节)上述假设是正确的。

当向量 s 的值初始化完成后，通过标准化处理，使得向量 s 中所有项相加和为1，以保证向量 s 为正确的状态初始分布向量。

式(11)和(12)为重启随机游走的过程：

$$\boldsymbol{r}^0=\boldsymbol{s}, \tag{11}$$

$$\boldsymbol{r}^{t+1}=(1-\lambda)\times\boldsymbol{T}\times\boldsymbol{r}^t+\lambda\times\boldsymbol{s}, \tag{12}$$

其中 $\boldsymbol{r}^t$ 表示重启随机游走的中间结果，t 表示迭代的次数，λ 为可调参数，取 λ=0.1。

令 $\boldsymbol{r}^{t+1}=\boldsymbol{r}^t$ 可求解最终平稳分布，如式(13)所示：

$$\boldsymbol{r}=\lambda(\boldsymbol{I}-c\boldsymbol{T})^{-1}\boldsymbol{s},\ c=1-\lambda。 \tag{13}$$

对于一个实体指称 m，求得其最佳目标实体，如式(14)所示：

$$m.e=\arg\max_{e}(\text{CP}(m,e)\times\boldsymbol{r}(e)), \tag{14}$$

其中，$m.e$ 表示实体指称 m 的最佳目标实体，$\boldsymbol{r}(e)$表示在平稳分布中候选实体 e 对应的值。

2.5 实体聚类

通过实体聚类，可以将 NIL 集合中表示同一种实体的实体指称聚为一簇。本文使用规则和 DBSCAN 聚类[19]相结合的算法对实体指称进行聚类，即先通过严格的规则对 NIL 集合中的实体指称进行粗粒度聚类，再通过聚类算法进行精确聚类。

对于两个实体指称 m_i 和 m_j，聚类规则如下：1)若 m_i 与 m_j 的表面字符串完全相同，m_i 与 m_j 的表面字符串与实体扩充之后的字符串完全相同，或两者实体扩充后的字符串完全相同，则将其暂时聚为一簇；2)若 m_i 与 m_j 实体扩充之后的字符串编辑距离小于 3，则将其暂时聚为一簇；3)若 m_i 与 m_j 实体扩充之后的字符串其中一个完全包含另一个字符串，则将其暂时聚为一簇；4)若 m_i 与 m_j 的命名实体类型不同，则不能将其聚为一簇。

利用上述规则对 NIL 集合中的实体指称进行粗粒度划分之后，使用 DBSCAN 聚类算法对每一个簇内部进行进一步划分。使用 LDA 主题模型将每个实体指称 m 对应的背景文本转换为向量表示，然后使用 DBSCAN 聚类算法进行聚类，得到最终结果。

3 实验

3.1 实验数据

本文在实体链接中使用的维基百科知识库为 2009 年 10 月份版本，总计 81 万余条实体条目，由 KBP 官方提供。实验部分使用 KBP的2014 年实体

链接官方评测数据集，总共包含 5234 条实体指称，其中在知识库中可以找到其目标实体的有 2817 条，在知识库中没有其对应的目标实体的有 2417 条。在所有实体指称中，1575 条来源于新闻文本，1743 条来源于网络文本，1916 条来源于论坛。

此外，为了进行更全面的对比，本文也使用了 Alhelbawy 等[9]以及 Han 等[10]的数据集，分别为 AIDA 数据集①和 IITB 数据集②。

3.2 评价方法

使用标准的 Wikification 评价方法对实体链接的性能进行评价，准确率、召回率和 F 值计算方法如式(15)~(17)所示：

$$P=\frac{\left|\mathrm{DP}_{\mathrm{expected}}\cap\mathrm{DP}_{\mathrm{actual}}\right|}{\left|\mathrm{DP}_{\mathrm{actual}}\right|},\tag{15}$$

$$R=\frac{\left|\mathrm{DP}_{\mathrm{expected}}\cap\mathrm{DP}_{\mathrm{actual}}\right|}{\left|\mathrm{DP}_{\mathrm{expected}}\right|},\tag{16}$$

$$F=\frac{2PR}{P+R},\tag{17}$$

其中，$\mathrm{DP}_{\mathrm{expected}}$ 表示实体链接系统的输出结果，$\mathrm{DP}_{\mathrm{actual}}$ 表示官方提供的标准答案。

对于实体聚类部分使用 CEAF 方法计算其准确率、召回率和 F 值[20]。

3.3 实验结果及分析

为验证本文方法的有效性，重现了 Alhelbawy 等[9]和 Han 等[10]的实验方法，其中对于待链接文本的预处理以及实体指称的扩充采用相同的处理方法，将本文实验结果与其对比，如表1所示。

由表 1 得出，本文提出的基于主题敏感的重启随机游走实体链接方法优于其他两种方法。由于 AIDA 数据集和IITB 数据集中不存在实体的聚类信息，因此没有对聚类部分的性能进行比较。

Alhelbawy 等[9]提出的基于 PageRank 的实体链接算法在计算状态转移概率矩阵时，只是将转移概率简单地根据图中节点的出度平均分配，并且图中仅包含候选实体 e 节点，不包含实体指称 m 节点。Han 等[10]的方法虽然对上述问题有所改进，但是在计算候选实体之间的相关度以及图的构建过程中存在不足。表 1 的实验结果对比说明，本文通过改进候选实体之间的相关度计算方法，完善了图的构建过程，并且通过确定随机游走过程中的重启节点以及计算在每个重启节点进行重启的概率，提升了整个实体链接系统的性能，证明此方法可行有效。

为了证明本文构建的实体名称多样性词表的作用，进行使用与不使用词表的对比实验，如表 2 所示。可以看出，基于LDA 主题模型构建的实体名称多样性词表在一定程度上发挥了积极作用。

本文通过对比实验，验证了主题敏感对整个实体链接系统的积极影响。通过在随机游走过程中设置每个候选实体 e 均为重启节点，将主题不敏感与主题敏感的方法进行对比，如表 3 所示。结果表明，基于主题敏感的重启随机游走包含实体指称的主题倾向信息，并且通过利用随机游走重启节点来增强此主题倾向信息，有利于提升整个实体链接系统的性能。

4 结束语

为了解决实体链接问题，本文提出一种基于主题敏感的重启随机游走实体链接方法。为了验证方法的有效性，在公开的数据集上进行实验，并与前人工作进行对比，实验结果证明了本文方法的有效性。本文方法仍有一些不足之处。例如，对于同一篇文本中的表面字符串相同的两个实体指称 m_i 和 m_j，若其表示的并非同一个实体，本文方法还不能处理这种情况。由于部分实体指称的候选实体数量

表 1　不同实体链接系统在不同数据集上的实现结果

Table 1　Effectiveness of various entity linking systems on different datasets

方法	KBP2014		AIDA	IITB
	Wikification-F 值	CEAF-F 值	Wikification-F 值	Wikification-F 值
Alhelbawy 等[9]	0.534	0.775	0.710	0.700
Han[10]	0.593	0.774	0.741	0.730
本文方法	0.623	0.799	0.762	0.751

① http://www.mpi-inf.mpg.de/yago-naga/aida/

② http://www.cse.iitb.ac.in/~soumen/doc/CSAW/

表 2 是否使用词表的实验结果对比

Table 2 Comparison of experimental results using the vocabulary table or not

实验名称	Wikification-*F* 值	CEAF-*F* 值
不使用词表	0.613	0.799
使用词表	0.623	0.799

表 3 是否使用主题信息的实验结果对比

Table 3 Comparison of experimental results using the topic information or not

实验名称	Wikification-*F* 值	CEAF-*F* 值
主题不敏感	0.619	0.799
主题敏感	0.623	0.799

庞大，整个实体链接系统的运行效率(在不影响性能的情况下)有待提高。

参考文献

[1] Roth D, Ji H, Chang M W, et al. Wikification and beyond: the challenges of entity and concept grounding // ACL 2014. Gaithersburg, MD, 2014: 7–18

[2] 谭咏梅，杨雪. 结合实体链接与实体聚类的命名实体消歧. 北京邮电大学学报, 2014, 37(5): 36–40

[3] Mihalcea R, Csomai A. Wikify!: Linking documents to encyclopedic knowledge. UNT Scholarly Works, 2007, 23(5): 233–242

[4] Cucerzan S. Large-scale named entity disambiguation based on Wikipedia data // Proc Joint Conference on Emnlp & Cnll. Prague, 2007: 708–716

[5] Dredze M, Mcnamee P, Rao D, et al. Entity disambiguation for knowledge base population // International Conference on Computational Linguistics. Beijing, 2010: 277–285

[6] Zheng Z, Li F, Huang M, et al. Learning to link entities with knowledge base // Proceedings of the Annual Conference of the North American Chapter of the Acl. Los Angeles, 2010: 483–491

[7] Pink G, Radford W, Cannings W, et al. Sydney CMCRC at TAC 2013 // Text Analysis Conference. Gaithersburg, MD, 2013: 1–6

[8] Dalton J, Dietz L. A neighborhood relevance model for entity linking // Open Research Areas in Information Retrieval. Lisbon, 2013: 149–156

[9] Alhelbawy A, Gaizauskas R. Graph ranking for collective named entity disambiguation // The 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014). Gaithersburg, MD, 2014: 75–80

[10] Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method // Proceedings of International Conference on Research & Development in Information Retrieval. Beijing, 2011: 765–774

[11] Pearson K. The problem of the random walk. Nature, 1905, 268: 2113–2122

[12] Grady L. Random walks for image segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(11): 1768–1783

[13] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the Web. Stanford Infolab, 1999, 9(1): 1–14

[14] Haveliwala T H. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge & Data Engineering, 2003, 15(4): 784–796

[15] Bradford R B. Use of latent semantic indexing to identify name variants in large data collections // Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on. Seattle, WA: IEEE, 2013: 27–32

[16] Sculley D. Web-scale k-means clustering // Proceedings of International Conference on World Wide Web. Raleigh, NC, 2010, 219: 1177–1178

[17] Milne D, Witten I H. Learning to link with wikipedia. Proceeding of ACM Conference on Information & Knowledge Management Norvig Peter Innovation in Search & Artificial Intelligence, 2008, 57(3): 509–518

[18] Cilibrasi R L, Vitanyi P M B. The Google Similarity Distance. Knowledge & Data Engineering IEEE Transactions on, 2007, 19(3): 370–383

[19] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of International Conference on Knowledge Discovery & Data Mining. Portland, OR, 1996: 226–231

[20] Luo X. On Coreference resolution performance metrics // Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Sydney: Association for Computational Linguistics, 2005: 25–32