

基于 NSGA-II 的大规模本体映射方法

薛醒思*

(福建工程学院 信息科学与工程学院, 福州 350108)

(*通信作者电子邮箱 jack8375@gmail.com)

摘要: 现有的基于进化算法的本体映射技术在面对大规模本体映射问题时, 由于搜索空间太大导致算法效率低下, 从而使其无法有效地在实际中得到应用。针对这一问题, 提出了基于快速非支配排序的多目标遗传算法 (NSGA-II) 的大规模本体映射方法。该方法通过三个步骤来映射本体: 1) 通过基于邻居相似度的划分算法来将源本体划分为不相交的概念块; 2) 通过相关概念过滤方法来确定目标本体中同源本体概念块相关的概念块; 3) 使用 NSGA-II 方法来完成概念块之间的映射并通过贪心算法集成最终的结果。使用 OAEI 2012 的小规模的书目本体测试数据集和大规模的生物医学本体测试数据集对所提出的方法进行验证。同 OAEI 2012 的参与者的比较结果表明, 所基于 NSGA-II 的大规模本体映射方法能够在较短的时间内获取较好的本体映射结果, 因此该方法是有效的。

关键词: 大规模本体映射; 本体划分算法; 快速非支配排序的多目标遗传算法

中图分类号: TP182 **文献标志码:** A

Large scale ontology aligning approach based on NSGA-II

XUE Xingsi*

(School of Information Science and Engineering, Fujian University of Technology, Fuzhou Fujian 350108, China)

Abstract: The application of existing ontology aligning technologies based on evolutionary algorithm is limited by the huge search space of large scale ontology aligning problem. To this end, in this paper, a large scale ontology aligning approach based on a fast elitist Non-dominated Sorting Genetic Algorithm for multi-objective optimization (NSGA-II) was proposed. To be specific, it worked in three steps: 1) a neighbor similarity based ontology partitioning algorithm was presented to split the source ontology into a set of disjoint concept blocks; 2) a relevant concept filtering method was proposed to determine the concept block in target ontology associated with each source one; 3) NSGA-II was utilized to align the various concept block pairs and a greedy algorithm was used to aggregate various results. Small scale bibliographic ontology benchmark and large scale biomedic ontology benchmark in OAEI 2012 were used to test the proposed approach. The comparisons with the participants of OAEI 2012 show that the large scale ontology aligning approach based on NSGA-II is able to determine good alignments in a short time, and therefore it is effective.

Key words: large scale ontology aligning; ontology partitioning algorithm; fast elitist Non-dominated Sorting Genetic Algorithm for multi-objective optimization (NSGA-II)

0 引言

本体被视为是语义网中数据异质问题的解决方法。然而由于人类的主观性, 同一个应用领域的不同的本体可能使用不同的方式来定义同一个实体, 使得本体间也存在着异质问题。本体间的异质问题是实现本体间协作的最大障碍, 解决这一问题需要确定不同本体中实体间的语义对应关系, 这一过程被称为本体映射。然而, 通过人为手动的方式来映射本体不仅耗时而且本体间的语义映射结果的正确性与完整性很难保证。因此, 近年来出现了大量的本体映射技术, 这些技术可以通过半自动化或自动化的方式来完成本体映射工作。其中, 基于进化算法的本体映射技术由于其能够获得较好的映射结果引起了广泛的关注^[1-5]。然而, 现有的基于进化算法的本体映射技术在面对大规模本体(拥有上百万个概念实体的大规模的本体)的映射问题时, 由于搜索空间太大导致算法效率低下, 从而使其无法有效地在实际中得到应用。此外, 对于大规模本体映射技术而言, 如何缩小待处理的数据规模

是找出正确实体映射的关键, 而将待映射的本体划分为小规模的分块, 通过映射相似的分块并集成多个映射结果是当前主流的技术。当前采用分块技术的大规模本体映射系统有 COMA++^[6]、Falcon-AO^[7]、Anchor-Flood^[8]、Lily^[9]、GOMMA^[10]、LogMAP^[11]等, 然而这些映射系统中采用的分块技术都没有考虑到分块过程中的映射目的, 即对本体执行分块的过程与后续的映射过程是相对独立的两个步骤, 这样就无法保证最终的本体映射结果的质量。

针对上述的问题, 本文提出了基于快速非支配排序的多目标遗传算法 (fast elitist Non-dominated Sorting Genetic Algorithm for multi-objective optimization, NSGA-II)^[12] 的大规模本体映射方法, 该方法采用了基于面向映射的分块技术来划分本体, 将分块间的映射过程视为一个多目标优化问题, 并通过多目标进化算法 NSGA-II 求解该问题。该方法的流程具体描述如下: 1) 通过基于邻近概念的自底向上的本体划分方法将源本体划分为一系列不相交的概念块; 2) 通过相关概念过滤的方法来确定同源本体概念块对应的目标本体概念块;

收稿日期: 2013-12-10; 修回日期: 2014-01-29。 基金项目: 福建省教育厅科研项目 (JA13227)。

作者简介: 薛醒思 (1981-), 男, 福建福州人, 讲师, 博士, 主要研究方向: 本体匹配、进化算法、数据挖掘。

3)通过NSGA-II方法求解源本体与目标本体概念块间的映射问题,并通过贪心算法集成最终的结果。

1 源本体划分

1.1 邻近概念相似度量

本文提出的本体划分算法采用的是邻近概念相似度量,这种度量比目前流行的本体划分算法——针对种类属性的鲁棒聚类算法 ROCK (ROBust Clustering using linKs)^[13]和针对网络的凝聚体系结构聚类算法 (Agglomerative Hierarchical Structural Clustering Algorithm for Networks, AHSCAN)^[14]采用的结构相似度量更为有效^[15]。

本体中不同概念块间的邻近概念相似度量取决于共同的邻近概念数量。给定两个概念块 C_1 和 C_2 , 它们的邻近概念相似度量由式(1)计算:

$$Sim_{neighbor}(C_1, C_2) = \frac{\sum_{i=1}^{|C_1|} \sum_{j=1}^{|C_2|} \frac{|NC(c_i) \cap NC(c_j)|}{|NC(c_i) \cup NC(c_j)|}}{|C_1| \cdot |C_2|} \quad (1)$$

其中: $|C_1|$ 和 $|C_2|$ 分别是两个概念块 C_1 和 C_2 中概念的个数, $|NC(c_i)|$ 是概念块 C_1 中某个概念 c_i 和它的邻近概念组成的集合, $|NC(c_j)|$ 是概念块 C_2 中某个概念 c_j 和它的邻近概念组成的集合。在本文的工作中,某个概念的邻近概念集合指的是在本体概念结构图中所有同该概念的最短路径距离小于等于2的概念。

此外,概念块集合 S 的内部相似度量由式(2)计算:

$$Sim_{intra}(S) = \frac{\sum_{i=1}^{|S|} Sim_{neighbor}(C_i, C_i)}{|S|} \quad (2)$$

其中: $|S|$ 是 S 中概念块的个数, C_i 表示 S 中第 i 个概念块。

1.2 源本体划分算法

本文提出的源本体划分算法由两个步骤组成:概念块初始化步骤和划分算法执行步骤。在第一步中,本体中的每一个概念独自构成一个概念块。在第二步中,算法根据概念块的大小和概念块之间的邻近概念相似度量来迭代地归并不同的概念块,形成一个临时概念块集合。如果形成的临时概念块集合中的概念块规模同当前概念块集合的规模相同,算法终止;否则,算法将临时概念块集合取代当前概念块集合,并继续归并更多的概念块。如果概念块的规模为1或生成的临时概念块集合的内部相似度量小于某个阈值 θ ,则算法终止。比起已有的本体划分算法而言,本文的本体划分算法在内存中只需存储一个概念块集合,在减少内存消耗的同时提高了算法的性能。

2 目标本体划分

目标本体的划分是在源本体概念块集合已经确定的前提下,通过相关概念过滤方法来实现的。该方法分为两个步骤进行:1)通过本体映射的字符串度量 (String Metric for Ontology Alignment, SMOA)^[16]比较目标本体和源本体概念块中不同概念的名称、标签和备注信息的相似程度,从目标本体中选取同源本体概念块对应的候选概念集;2)通过计算候选概念集中每个概念同源本体概念块间的关联值来进一步确定目标本体同源本体概念块对应的相关概念集。

在第1)步中,对于目标本体 O_{tgt} 中的每一个概念 c_i , c_i 同 Seg_{src} 之间的相似值 s_i 等于 c_i 同 S_{src} 中每一个概念 c_j 的相似值 s_{ij} 之和 (只对大于阈值 α 的 s_{ij} 求和)。如果得到的 s_i 大于阈

值 β ,则将 c_i 加入候选概念集 C 。

在第2)步中需要进一步确定候选概念集中的同源本体概念块的相关概念以进一步减小后续本体映射过程中的搜索空间。首先,在第一步获取的候选概念集中,概念 c_k 对概念 c_i 的影响值定义如下:

$$influ_k(c_i) = s_k \times e^{-dist(c_k, c_i)^2} \quad (3)$$

其中: s_k 表示概念 c_k 同源本体概念块集合之间的相似值; $dist(c_k, c_i)$ 表示概念 c_k 同概念 c_i 在本体概念图中的最短路径的长度,影响值随着 $dist(c_k, c_i)$ 的增加而减少。接下来,通过式(4)计算概念 c_i 同源本体概念块的关联值:

$$relevant(c_i) = s_i \times influ(c_i) \quad (4)$$

其中: $influ(c_i) = \sum influ_k(c_i)$, $dist(c_k, c_i) \leq 2$ 。如果 c_i 的关联值大于阈值 γ ,则将 c_i 加入到目标本体概念块中。

3 使用NSGA-II映射本体概念块

3.1 本体映射结果评价

在本体映射领域中,通常使用源自信息检索领域的两个质量衡量指标 (即:查全率 P_{recall} 和查准率 $P_{precision}$) 来评价本体映射结果。但是使用这两个指标工作的前提是必须要有专家给出的标准本体映射结果,然而这一结果在实际应用中是不存在的。为了克服这一缺陷,在最终的本体映射结果是1:1的前提下,本文提出了两种指标 (即:匹配覆盖率 $P_{MatchCoverage}$ 和匹配重复率 $P_{Frequency}$) 来分别近似地获取映射结果的查全率和查准率。

给定两个本体概念块 S_1 和 S_2 , $P_{MatchCoverage}$ 和 $P_{Frequency}$ 分别由以下两个公式计算:

$$P_{MatchCoverage} = \frac{|E_{S_1-Match}| + |E_{S_2-Match}|}{|E_{S_1}| + |E_{S_2}|} \in [0, 1] \quad (5)$$

$$P_{Frequency} = \frac{|E_{S_1-Match}| + |E_{S_2-Match}|}{2|Corr_{s_1-s_2}|} \in [0, 1] \quad (6)$$

其中: $|E_{S_1-Match}|$ 和 $|E_{S_2-Match}|$ 分别是 S_1 和 S_2 中映射上的实体 (概念块中的概念、概念的属性和概念的实例统称实体) 个数, $|E_{S_1}|$ 和 $|E_{S_2}|$ 分别是 S_1 和 S_2 中所有实体的个数, $|Corr_{s_1-s_2}|$ 是映射结果中的映射个数。 $P_{MatchCoverage} = 1$, 意味着映射结果的查全率很高;同样地, $P_{Frequency} = 1$, 意味着映射结果的查准率很高。

3.2 本体概念块映射问题的优化模型

给定两个本体概念块 S_1 和 S_2 , 本体概念块映射问题的多目标优化模型如下:

$$\max f(X) = \max(P_{MatchCoverage}(X), P_{Frequency}(X)) \quad (7)$$

$$\text{s. t. } X = (x_1, x_2, \dots, x_n)^T$$

$$x_i \in [0, |entitySet_{s_2}|]; i = 1, 2, \dots, n$$

其中: $n = |entitySet_{s_1}|$, $|entitySet_{s_1}|$ 和 $|entitySet_{s_2}|$ 分别表示 S_1 和 S_2 中实体的个数。该模型的目标是同时最大化 $P_{MatchCoverage}$ 和 $P_{Frequency}$ 值。

本文使用NSGA-II来求解该优化问题。NSGA-II是灵活的、鲁棒性强的优化算法,该算法能快速找到多目标优化问题中的各种非支配解。该算法首先对当前的种群使用标准的交叉与变异算子,然后通过快速非支配排序技术与拥挤度距离来产生下一代群体,最后兼顾了非支配性与多样性的最优个体被选为多目标优化问题的解集。NSGA-II算法中的4个基本步骤如下。

3.2.1 编码机制

在本文的工作中,个体编码信息既包括用于集成不同相

似度度量的映射结果的权重,也包括用于过滤本体映射结果的阈值。本文采用的是加权平均的方法集成不同的相似度量产生的映射结果,具体描述如下:

$$\varphi(s(c), w) = \sum_{i=1}^n w_i s_i(c) \quad (8)$$

其中: $\sum_{i=1}^n w_i = 1$, $w_i \in [0, 1]$; $s(c)$ 是不同的相似度量获得的映射结果向量; w 是权重向量; n 是相似度度量的个数。考虑到权重的特点,本文的编码通过在区间 $[0, 1]$ 内定义分割点来间接地表示不同的权重。假设 p 是所需的权重个数,则分割点集合可以表示为 $c' = \{c'_1, c'_2, \dots, c'_{p-1}\}$ 。译码过程分为两个步骤:1) 将分割点集合中的元素按照升序排列,得到新的集合 $c = \{c_1, c_2, \dots, c_{p-1}\}$; 2) 按照式(9) 计算不同的权重:

$$w_k = \begin{cases} c_1, & k = 1 \\ c_k - c_{k-1}, & 1 < k < p \\ 1 - c_{p-1}, & k = p \end{cases} \quad (9)$$

用于过滤本体映射结果的阈值用一位编码表示,其取值范围是 $[0, 1]$ 。

3.2.2 适应度函数

适应度函数是用于评价通过个体编码中的权重和阈值获取的本体映射结果质量的目标函数。本文采用两个目标函数,分别用于计算 $P_{MatchCoverage}$ 和 $P_{Frequency}$ 。

3.2.3 遗传算子

1) 选择算子。本文采用的选择算子首先根据群体中不同个体的拥挤度进行降序排序,并选择排在前半部分的个体,从中随机复制一个个体直到形成新的群体。

2) 交叉算子。本文采用的是单点交叉算子。首先在两个个体中随机确定一个分割点,该分割点将两个父个体分割为左右两部分。然后通过交换两个父个体右边部分的编码以产生新的两个子个体。

3) 变异算子。本文采用的是位点变异算子。首先根据变异概率确定对个体会产生变异的编码位,然后将这些编码位的值从 1 修改为 0,或是从 0 修改为 1。

3.2.4 生成下一代个体

本文首先通过将当前代种群与新生成的种群放在一起,消除冗余的个体。通过快速非支配排序算法^[12]并根据不同个体间的拥挤度来选出新的群体。

当算法终止后,从 Pareto 前沿中选出 3 个拐点解作为代表。由于 Pareto 前沿的拐点区域代表了 Pareto 前沿中不同目标间的最大权衡,在 Pareto 前沿的拐点区域中的解有以下特点:在一个目标上的小改进会导致在至少一个的其他目标上较大的恶化。在没有用户偏好信息的前提下, Pareto 前沿的拐点区域中的解被默认是决策制定者需要的解^[17]。本文选择的 3 个拐点解分别是拥有最好的 $P_{MatchCoverage}$ 、 $P_{Frequency}$ 以及二者间最好的权衡。具体地说,在 Pareto 前沿中拥有最好 $P_{MatchCoverage}$ 的解中选出一个 $P_{Frequency}$ 最高的解。同样地,在 Pareto 前沿中拥有最好 $P_{Frequency}$ 的解中选出一个 $P_{MatchCoverage}$ 最高的解。关于 $P_{MatchCoverage}$ 和 $P_{Frequency}$ 间最好的权衡的解,通过二者的和谐均值 $MatchFmeasure$ 来度量:

$$MatchFmeasure = \frac{2 \times P_{MatchCoverage} \times P_{Frequency}}{P_{MatchCoverage} + P_{Frequency}} \quad (10)$$

在 Pareto 前沿中拥有最好的 $MatchFmeasure$ 的解被选为第 3 个代表解。

最后,通过 NSGA-II 方法获取的不同概念块之间的映射结果经过贪心算法集成,以获取最终的本体映射结果。

4 实验结果与分析

本文采用的是由 OAEI 2012^[18] 提供的 Bibliographic、Anatomy 和 Library 测试数据集。接下来会给出本文方法的参数配置,并将结果同 OAEI 2012 的参与者进行比较。表 1~2 给出了小规模的书目本体测试数据集和大规模生物医学本体测试数据集的结果,其中 OAEI 2012 的参与者的结果源自参考文献^[19],符号 R 、 P 和 F 分别代表本体映射结果的 P_{recall} 、 $P_{precision}$ 和二者的和谐均值 F 度量 (F -measure)。本文方法的结果是在算法 30 次独立运行中的均值。

4.1 实验配置

本体划分算法的参数如下:

- 1) 概念块相似度下限 $\delta = 0.8$;
- 2) 概念块中概念数量上限 $\varepsilon = 60$;
- 3) 概念块集合的内部相似度下限 $\theta = 0.2$ 。

相关概念过滤方法中的参数如下:

- 1) 目标本体概念与源本体概念之间的相似度下限 $\alpha = 0.6$;
- 2) 目标本体概念与源本体概念块之间的相似度下限 $\beta = 0.7$;
- 3) 目标本体概念与源本体概念块之间的相关度下限 $\gamma = 0.1$ 。

本体划分算法和相关概念过滤方法的参数是由实验确定的。在划分算法中,希望产生的概念块集合以及每个概念块中的概念个数不应当太大,这样可以提高后续过程的处理效率。在相关概念过滤方法中,产生的目标本体相关概念块的规模不应当太大,而真正同源本体概念块相关的概念不应当被过滤掉,这样可以提高后续映射过程的效率。

NSGA-II 算法采用以下的配置:

- 1) 采用的概念间相似度量技术分别是:SMOA 度量、基于 WordNet 的度量^[20] 和基于 profile 的度量^[21];
- 2) 每个参数的搜索空间是连续的区间 $[0, 1]$;
- 3) 数值精度为 0.01;
- 4) 群体规模为 20 个个体;
- 5) 交叉概率为 0.8;
- 6) 变异概率为 0.09;
- 7) 最大进化代数 5 代。

运行算法的硬件配置如下:

- 1) 处理器为 Xeon 5472 (4 核);
- 2) CPU 速度为 3 GHz;
- 3) RAM 容量为 8 GB。

4.2 小规模的书目本体测试数据集

书目测试数据集由一组小规模的本体组成,这些本体都是由源本体变化生成的。源本体中有 33 个实名类,24 个对象属性,40 个数据属性,56 个实名个体和 20 个匿名个体。总的来说,书目测试数据集分为三组:1) 简单测试数据集 ($1 \times \times$) 将目标本体同自身进行映射;2) 系统测试数据集通过改变源本体中的一些特征,例如实体的名称、备注、体系结构等来生成目标本体 ($2 \times \times$);3) 来源于 Web 的本体 ($3 \times \times$)。

表 1 中给出了不同系统在测试数据集 $1 \times \times$ 、 $2 \times \times$ 和 $3 \times \times$ 上的运行结果的评价均值,从表 1 可看出:本文方法获取的本

体映射结果的 F -measure 值优于所有 OAEI 2012 的参与者,然而本文方法的运行时间为 15 s(其中步骤 1 用时 7 s,步骤 2 用

时 2 s,步骤 3 用时 6 s),排名第 4 位。因此,对于小规模的本体映射问题而言,本文方法是有效的。

表 1 本文方法同 OAEI 2012 参与者在书目测试数据集上的比较

系统	R	P	F	运行时间/s	系统	R	P	F	运行时间/s
MapSSS	0.77	0.99	0.87	35	WikiMatch	0.54	0.74	0.62	750
YAM++	0.72	0.98	0.83	120	ServOMap	0.43	0.88	0.58	18
AROMA	0.64	0.98	0.77	8	LogMap	0.45	0.73	0.56	28
AUTOMSV2	0.54	0.97	0.69	80	MaasMatch	0.57	0.54	0.56	38
WeSeE	0.53	0.99	0.69	650	MEDLEY	0.50	0.60	0.54	85
Hertuda	0.54	0.90	0.68	9	ServOMapLt	0.20	1.00	0.33	7
HotMatch	0.50	0.96	0.66	20	ASE	0.54	0.49	0.51	40
Optima	0.49	0.89	0.63	380	本文方法	0.82	0.95	0.88	15

4.3 大规模的生物医学本体测试数据集

大规模的生物医学本体测试数据集采用 3 个分别拥有 78 989,306 591 和 66 724 个概念类的大规模的生物医学本体 FMA、SNOMED CT 和 NCI。该测试数据集的任务分为 3 个本体映射子任务:任务 1(FMA-NCI)、任务 2(FMA-SNOMED)和任务 3(SNOMED-NCI),其中每一个映射任务使用到不同的输入本体。

从表 2 可看出:本文方法在任务 1 中得到的映射结果的 F -measure 值在 OAEI 2012 所有参与者中排名第 2,运行时间

为 230 s(其中步骤 1 用时 68 s,步骤 2 用时 136 s,步骤 3 用时 26 s),排名第 6 位。在任务 2 中,本文方法得到的映射结果的 F -measure 值比所有 OAEI 2012 参与者的结果都要好,运行时间为 674 s(其中步骤 1 用时 68 s,步骤 2 用时 136 s,步骤 3 用时 470 s),排名第 4 位。在任务 3 中,本文方法得到的映射结果的 F -measure 值在所有 OAEI 2012 参与者中排名第 4,运行时间为 1 355 s(其中步骤 1 用时 315 s,步骤 2 用时 542 s,步骤 3 用时 498 s),排名第 5 位。因此,总的来说,对于大规模的本体匹配问题而言,本文方法是有效的。

表 2 本文方法同 OAEI 2012 参与者在生物医学测试数据集上的比较

系统	任务 1: FMA-NCI				任务 2: FMA-SNOMED				任务 3: SNOMED-NCI			
	R	P	F	运行时间/s	R	P	F	运行时间/s	R	P	F	运行时间/s
YAM++	0.83	0.85	0.84	217	0.68	0.81	0.74	23900	0.60	0.79	0.68	30 155
ServOMapL	0.81	0.87	0.84	251	0.69	0.88	0.77	517	0.59	0.79	0.68	738
ServOMap	0.76	0.89	0.82	104	0.65	0.87	0.75	532	0.56	0.83	0.67	654
LogMap-noe	0.79	0.87	0.83	206	0.62	0.82	0.71	791	0.57	0.81	0.67	1 505
LogMapLt	0.81	0.68	0.74	55	0.18	0.85	0.30	717	0.56	0.74	0.63	178
LogMap	0.78	0.86	0.82	131	0.62	0.83	0.71	612	0.57	0.81	0.67	955
GOMMA-bk	0.87	0.80	0.83	231	0.86	0.56	0.68	1 893	0.61	0.66	0.64	1 940
GOMMA	0.83	0.89	0.86	1 304	0.24	0.35	0.29	1 994	0.53	0.71	0.61	1 820
本文方法	0.86	0.87	0.85	230	0.72	0.85	0.78	674	0.58	0.79	0.67	1 355

5 结语

本文提出了基于 NSGA-II 的大规模本体映射方法,该方法通过 3 个步骤来映射本体:1)通过自底向上的基于邻居相似度的划分算法来将源本体划分为不相交的概念块;2)通过相关概念过滤方法来确定目标本体中同源本体概念块相关的概念块;3)使用 NSGA-II 方法来完成概念块之间的映射。本文使用 OAEI 2012 的书目测试数据集和生物医学测试数据集的对本文方法进行测试,同 OAEI 2012 参与者的比较结果表明本文方法能够在较短的时间内获取较好的本体映射结果。

参考文献:

- [1] MARTINEZ-GIL J, ALBA E, ALDANA-MONTES J F. Optimizing ontology alignments by using genetic algorithms [C]// Proceedings of the 2008 Workshop on Nature based Reasoning for the Semantic Web. Karlsruhe: [s. n.], 2008: 21–45.
- [2] NAYA J M V, ROMERO M M, LOUREIRO J P, et al. Improving ontology alignment through genetic algorithms [M]// Soft Computing Methods for Practical Environment Solutions: Techniques and Studies. Coruna: University of A Coruna, 2010: 240–259.
- [3] ALEXANDRU-LUCIAN G, IFTENE A. Using a genetic algorithm for optimizing the similarity aggregation step in the process of onto-

gy alignment [C]// Proceedings of the 2010 9th Roedunet International Conference. Piscataway: IEEE Press, 2010: 118–122.

- [4] WANG J, DING Z, JIANG C. GAOM: genetic algorithm based ontology matching [C]// APSCC 2006: Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing. Piscataway: IEEE Press, 2006: 617–620.
- [5] BOCK J, HETTENHAUSEN J. Discrete particle swarm optimisation for ontology alignment [J]. Information Sciences, 2012, 192: 152–173.
- [6] ALGERGAWY A, MASSMANN S, RAHM E. A clustering-based approach for large-scale ontology matching [C]// ADBIS 2011: Proceedings of the 15th international Conference on Advances in Databases and Information Systems. Berlin: Springer-Verlag, 2011: 415–428.
- [7] HAMDI F, SAFAR B, REYNAUD C, et al. Alignment-based partitioning of large-scale ontologies [C]// Advances in Knowledge Discovery and Management, Studies in Computational Intelligence 292. Berlin: Springer-Verlag, 2010: 251–269.
- [8] SEDDIQUI M H, AONO M. An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(4): 344–356.

(下转第 1630 页)

- 241.
- [4] MEI Q, LING X, WONDRA M, *et al.* Topic sentiment mixture: modeling facets and opinions in weblogs [C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM Press, 2007: 171 – 180.
 - [5] CANDÈS E J, LI X, MA Y, *et al.* Robust principal component analysis? [J]. *Journal of the ACM*, 2011, 58(3): 11.
 - [6] MIN K, ZHANG Z, WRIGHT J, *et al.* Decomposing background topics from keywords by principal component pursuit [C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2010: 269 – 278.
 - [7] LANDAUER T K, FOLTZ P W, LAHAM D. An introduction to latent semantic analysis [J]. *Discourse Processes*, 1998, 25(2/3): 259 – 284.
 - [8] LIN Z, CHEN M, MA Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices [EB/OL]. [2013-08-16]. <http://arxiv.org/pdf/1009.5055v3.pdf>.
 - [9] BLEI D M, LAFFERTY J D. Dynamic topic models [C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006: 113 – 120.
 - [10] BLEI D M, LAFFERTY J D. A correlated topic model of science [J]. *The Annals of Applied Statistics*, 2007, 1(1): 17 – 35.
 - [11] BLEI D M, McAULIFFE J D. Supervised topic models [EB/OL]. [2013-08-16]. <https://papers.nips.cc/paper/3328-supervised-topic-models.pdf>.
 - [12] BRANAVAN S R K, CHEN H, EISENSTEIN J, *et al.* Learning document-level semantic properties from free-text annotations [J]. *Journal of Artificial Intelligence Research*, 2009, 34: 173 – 196.
 - [13] BRODY S, ELHADAD N. An unsupervised aspect-sentiment model for online reviews [C]// HLT 2010: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 804 – 812.
 - [14] CHANG J, BLEI D M. Relational topic models for document networks [C]// Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Cambridge: MIT Press, 2009: 81 – 88.
 - [15] CHEMUDUGUNTA C, SMYTH P, STEYVERS M. Combining concept hierarchies and statistical topic models [C]// CIKM 2008: Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2008: 1469 – 1470.
 - [16] FUNG G P C, YU J X, YU P S, *et al.* Parameter free bursty events detection in text streams [C]// VLDB 2005: Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim: VLDB Endowment, 2005: 181 – 192.
 - [17] ALSUMAIT L, BARBARA D, DOMENICONI C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking [C]// ICDM 2008: Proceedings of the 8th IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2008: 3 – 12.
 - [18] JO Y, OH A H. Aspect and sentiment unification model for online review analysis [C]// WSDM 2011: Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 815 – 824.
 - [19] SAMMUR, WEBB G I. Encyclopedia of machine learning [M]. New York: Springer-Verlag, 2010: 280 – 287.
 - [20] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. *Machine Learning*, 2001, 42(1/2): 177 – 196.

(上接第1625页)

- [9] WANG P, XU B. An effective similarity propagation method for matching ontologies without sufficient or regular linguistic information [C]// Proceedings of the ASWC 2009, LNCS 5926. Berlin: Springer-Verlag, 2009: 105 – 119.
- [10] KIRSTEN T, GROSS A, HARTUNG M, *et al.* GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution [J]. *Journal of Biomedical Semantics*, 2011, 2: 6.
- [11] JIMÉNEZ-RUIZ E, GRAU B C. LogMap: logic-based and scalable ontology matching [C]// Proceedings of the Semantic Web — ISWC 2011, LNCS 7031. Berlin: Springer-Verlag, 2011: 273 – 288.
- [12] DEB K, AGRAWAL S, PRATAP A, *et al.* A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II [C]// Proceedings of Parallel Problem Solving from Nature PPSN VI, LNCS 1917. Berlin: Springer-Verlag, 2000, 1917: 849 – 858.
- [13] GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes [J]. *Information Systems*, 2000, 25(5): 345 – 366.
- [14] YURUK N, METE M, XU X, *et al.* AHSCAN: agglomerative hierarchical structural clustering algorithm for networks [C]// ASONAM 2009: Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining. Washington, DC: IEEE Computer Society, 2009: 72 – 77.
- [15] SARULADHA K, AGHILA G, SATHIYA B. A partitioning algorithm for large scale ontologies [C]// Proceedings of the 2012 International Conference on Recent Trends in Information Technology. Piscataway: IEEE Press, 2012: 526 – 530.
- [16] STOILLOS G, STAMOU G, KOLLIAS S. A string metric for ontology alignment [C]// Proceedings of the Semantic Web — ISWC 2005. Berlin: Springer-Verlag, 2005: 624 – 637.
- [17] BECHIKH S, SAID L B, GHÉDIRA K. Searching for knee regions of the Pareto front using mobile reference points [J]. *Soft Computing*, 2011, 15(9): 1807 – 1823.
- [18] Ontology Alignment Evaluation Initiative (OAEI) [EB/OL]. [2013-09-28] <http://oaei.ontologymatching.org/2012/>.
- [19] AGUIRRE J L, GRAU B C, ECKERT K, *et al.* Results of the ontology alignment evaluation initiative 2012 [EB/OL]. [2013-10-10]. <http://oaei.ontologymatching.org/2012/results/oaei2012.pdf>.
- [20] MILLER G A. WordNet: a lexical database for English [J]. *Communications of the ACM*, 1995, 38(11): 39 – 41.
- [21] MAO M, PENG Y, SPRING M. An adaptive ontology mapping approach with neural network based constraint satisfaction [J]. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2010, 8(1): 14 – 25.