DOI：10.16353/j.cnki.1000-7490.2015.10.028

● 柏　晗，成　颖，柯　青（南京大学　信息管理学院，江苏　南京　210023）

网络检索结果聚类研究综述*

摘　要： 网络检索结果聚类是将搜索引擎的检索结果聚类为有意义的类别，并赋予标签描述，以便用户快速获得所需信息的技术。文章根据网络检索结果聚类算法的改进方向将其分为面向经典和面向标签的聚类算法两类。前者的改进主要有优化特征选择、优化聚类数 K 以及生成重叠聚类等；后者的改进主要有优化类计分运算、优化类合并运算、数据结构优化、候选标签选择以及基于语义的优化等。在对相关研究进行综述的基础上探讨了检索结果聚类面临的问题和未来的发展方向。

关键词： 检索结果聚类；检索算法；优化；综述

Abstract: By clustering search results into meaningful clusters and giving the appropriate description labels, network search results clustering is used to help to get the information quickly. Based on the improving direction of network search results clustering algorithm, this paper divides the algorithm into classical clustering oriented algorithm and label-oriented algorithm. The improvement of classical clustering oriented algorithm mainly includes optimization of feature selection, optimization of cluster number K and generation of overlapping clustering. The improvement of label-oriented algorithm mainly includes optimization of class scoring operation, optimization of class merging operation, optimization of data structure, selection of candidate label and optimization based on semantics. On the basis of the review of relevant studies, the paper discusses the existing problems and future developing directions of search results clustering.

Keywords: search results clustering; retrieval algorithm; optimization; survey

面对互联网上的海量信息，用户借助搜索引擎获得的结果中只有极少部分与需求相关。对此，学界提出了不同的改进策略。一是检索结果的多样化（Diversification）[1]。多样化不仅考虑文档与查询的相关性，还将文档间的不相似性作为排序的指标，试图从检索文档集层面优化检索结果的排序[2]。二是检索结果聚类（Search Results Clustering）[3]。用户可以根据聚类标签直接定位感兴趣的结果，还可以根据其他类别的标签更好地了解查询词，必要时可以重构检索策略。

检索结果聚类与文本聚类的共性是都需要考虑聚类的相关性以及生成标签的质量，不同点在于前者还需考虑重叠聚类（软聚类）、速度以及片段容忍[4]（即基于信息较少的网页片段 snippet 完成聚类）。有关聚类方法已经有 Xu Rui[5] 以及 Rokach[6] 等学者完成的系统综述，不过，这些工作主要涉及算法本身及其衍化，甚少关注网络检索结果短文本聚类的特定语境。Carpineto[7] 等对检索结果聚类搜索引擎进行了概貌性介绍，对检索结果聚类算法及其改进着墨不多。鉴于检索结果聚类已经形成的丰富研究成果，对其进行系统梳理尤显必要。

1　检索结果聚类

检索结果聚类可分为片段获取、预处理、特征选择、标签生成和聚类结果展示 5 个阶段。其中预处理与其他自然处理应用没有差别，不再赘述。针对片段获取，Zamir 等[4] 对比了网页文档和网页片段的聚类结果，结果显示片段中包含有助于聚类的项，同时去除了原始文档中可能导致错误分类的“噪音”。片段相较于原始文档约损失 15% 的聚类准确率。聚类结果展示主要有 3 种结构，分别是扁平划分、层次结构（通常以树的形式展示）以及图。

根据侧重点的不同，聚类算法可分为以数据为中心和以描述为中心[7]两类。前者更加注重算法，代表性的工作是 1992 年基于经典聚类算法 K-means 的 Scatter/Gather 系统，该系统有无法产生重叠类、聚类标签的可理解性较差、不能很好地适应网络聚类等不足。后者则更加关注聚类结果的描述（聚类标签），典型的例子有 1999 年 Zamir 等提出的后缀树聚类算法（Suffix Tree Clustering，STC）[3-4] 和 lingo 算法[8]，基本思想是基于高频短语的共

* 本文受国家社会科学基金重大招标项目“面向学科领域的网络信息资源深度聚合与服务研究”（项目编号：12&ZD221）和中国科学技术信息研究所“大数据环境下的人机交互研究”课题的资助。

现，算法通过提取信息量大的短语作为聚类依据，最终的标签也从短语中产生。在这两项经典工作的基础上，目前的检索结果聚类算法主要有两个发展方向：一是对经典的文本聚类算法进行合理的改进以使之适应检索结果聚类的要求；另一则是根据网页片段的特性，对包含 STC 等基于标签的算法进行改进，图 1 是两类算法改进的总体概貌。

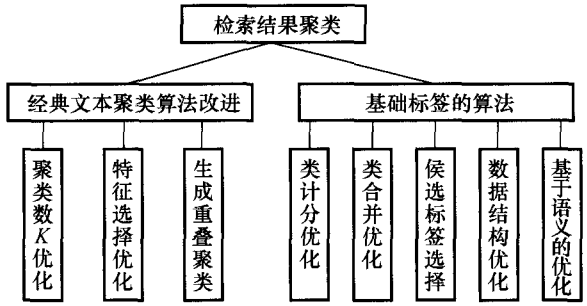

图 1　检索结果聚类算法改进的总体概貌

1.1　经典聚类算法的改进

该领域的工作是指采用经典聚类方法，如 K-means，AHC 等进行检索结果聚类研究[9]。有关经典聚类方法的主要科学内涵特征、技术能力指标和应用指标等请参阅 Xu Rui[5] 以及 Rokach[6] 等工作，本文主要阐述研究者在检索结果聚类中的相关改进工作。

1.1.1　软聚类　多数经典的聚类算法属于硬聚类，无法产生重叠聚类（软聚类），而实际应用中存在一个网络文档包含多个主题的情况，所以软聚类的实现是经典算法需要面对的问题。

1）相似度阈值。Wang 等[9]在 K-means 算法中设定了相似度阈值，当片段与类簇中心的相似度超过该阈值时则片段归入该类，从而产生重叠类。Maiti 等[10]整合了 K-means 以及分裂式层次算法，利用后者确定初始聚类中心，然后通过设定相似度阈值，基于 K-means 算法完成聚类。

2）最小化目标函数。Wang Fei 等[11]基于最小化目标函数（公式 1）在模糊 c-means（FCM）算法中实现了重叠分类。

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \| x_i - V_j \|^2, 1 \leqslant m \leqslant \infty \qquad (1)$$

式中，m 是权重指数，m 越大，模糊划分越显著；u_{ij}是片段 i 对类 j 的隶属度；v_j 是第 j 个类簇的中心；$\|x_i - v_i\|$是片段 i 和类 j 中心的距离（相似度）。通过迭代优化 u_{ij}和 v_j（公式 2）。

$$u_{ij} = 1 / \sum_{k=1}^{c} \left(\|x_i - v_j\| / \|x_j - v_k\| \right)^{2/(m-1)}$$
$$v_j = \sum_{i=1}^{N} u_{ij}^m * x_i / \sum_{i=1}^{N} u_{ij}^m \qquad (2)$$

直到满足公式（3），其中 k 为迭代次数，ε 是介于 0 到 1 的终止条件。

$$\max_{ij} \{ | u_{ij}^{k+1} - u_{ij}^{k} | \} < \varepsilon \qquad (3)$$

1.1.2　特征选择优化

1）高频短语。以项为特征的聚类效果往往难以满足需求。文献［3］经试验证明高频短语比项聚类效果更佳，准确率可以提升 20%。

2）共现信息。Navigelli 等[12-13]提出了将词义归纳（Word Sense Induction，WSI）应用于检索结果聚类的思路。WSI 即在粗语料中进行词义自动发现，使用基于图的算法在用户查询的共现图中计算出最大生成树以识别出查询词的语义，然后完成聚类。基于 Google Web1T 语料库的测试表明，该算法的准确率高达 85.24%，而 STC 算法则仅为 54.29%。Sha 等[14]的工作思想与 Navigelli[12] 有相似之处，即以片段集中的项 w 为节点集，定义 $S_{ij} = \frac{2 * |D_i \cap D_j|}{|D_i| + |D_j|}$为边 $w_i w_j$ 的标注，定义与 w_i 共现的项数量作为 w_i 的度（Degree），通过不同项的度和边的相似性发现项关系并进行节点的合并以形成类簇，该算法平均 F 值为 0.7，而 STC 和 K-means 算法则在 0.6 和 0.4 左右。Zhang 等[15]提出了 CoHC 算法，该算法首先在语料库中检索项的 2-gram 共现信息；将 2-gram 组合为 n-gram；在清除冗余的 n-gram 之后，将剩下的 n-gram 进行排序并将其作为聚类标签和基类；采用层次聚类算法完成聚类。在 ANMI 指标上，CoHC 比 Vivisimo 至少超出 8.87%，比 STC 则至少超出 41.31%。

1）链接信息。在信息检索研究中，网页间的链接关系能够提供有价值的信息[16]。Wang 等[9]将网页的出链和入链作为特征，以链接的共引和共被引为特征进行聚类，成功地将基于链接的特征用于聚类算法。随后，Wang 等[17]又提出了一种结合网页间的共引链接信息、内容信息以及锚文本信息的结果聚类算法，对比基于项的聚类，准确率从 0.8 提升到 0.9，召回率则从 0.85 提升到 0.88。

2）词汇相关度（Lexical Affinity，LA）。LA 的优点是结构上比短语更加灵活，能够发现并不临近的相关项。Lassi[18]是一个基于 AHC 的检索结果聚类系统，以文本中成对出现的 LA 代替项，结果表明聚类效果可以提高 30%。

3）外源性知识。Hu 等[19]利用 Wikipedia 和 WordNet 的背景知识作为外部特征，层次化短语作为内部特征，确定聚类中心，结果表明改进后的 K-means 算法 F 值可提高 30.39%，平均准确度提高 7.83%。

4）多特征融合。张刚等[20]针对传统聚类算法难以生成有意义类别标签的问题提出了将 DF、查询日志、查询

词上下文等特征融合的类别标签抽取算法，采用基于标签的 GBCA 算法完成聚类，实验结果表明：特征融合优于单一特征，在采用 3 类特征时效果最优；标签抽取效果与 STC 相比，在 $P@3$ 和 $P@5$ 指标上，分别提高 58.62% 和 42.4%，在 F-Measure 指标上，则比 STC 提高 0.1 以上。

1.1.3 优化聚类数 K ①设定阈值。K-means 等算法需要预先设定聚类数 K，层次聚类算法也需要设定终止条件。网页内容的差异使得初始 K 值难以确定，K-means 等算法的基本思想是将文档归入与聚类中心最近的类，如果对该距离设定阈值，小于该阈值则产生新类，即可避免 K 的预先设定[9]。②AP 算法。为解决 FCM 类簇数的选择问题，Wang Fei[11]引入了 AP（Affinity Propagation）算法，即先设定一个较大的初始聚类数，通过 AP 算法优化类的数量，通过迭代运算获得最终的聚类数。③蚁群算法。Schockaert 等[21]提出的基于蚁群思想的检索结果聚类算法具有无需事先设定聚类数以及无需初始化的优点，并结合模糊逻辑对其进行了改进。④链接信息。夏斌等[22]利用链接信息发现多个权威网页作为初始聚类中心，在避免确定 K 值的同时也提高了聚类的准确性。⑤检索结果数量。针对新闻检索结果聚类这一特定领域，Cheng Jia[23]提出了独特的 K 值确定方法（公式 4）。其中 N 为 Google 新闻检索结果数量，$|C_i|$ 表示类簇 C_i 中项的数量。

$$K = N\frac{\left|\bigcup_{i=1}^{N} C_i\right|}{\sum_{i=1}^{N} C_i} \tag{4}$$

1.1.4 一点思考 从经典文本聚类算法在检索结果聚类中的改进以及应用中可以看出。第一，在对经典算法进行改进的路径上，特征选择优化较为成功，共现信息、链接信息、高频短语、外源性知识、词汇相关度及其融合应用都取得了显性的聚类效果。在后继的研究与应用中，选取项之外更具语义的特征将是自然语言处理领域的一个共性课题，也是研究的一个重要方向。第二，虽然可以通过设置相似度阈值以及最小化目标函数实现软聚类，但是随之产生了新的难题，即如何设置阈值，最小化目标函数中的 ε 以及相似度阈值在实践中如何找到最优解远非易事。第三，在优化 K 的工作中，各项研究的技术思路差异较大，Cheng Jia[23]的工作没有说明公式（4）的具体依据，因此其内在的合理性还有待探讨；夏斌等[22]在论文中没有进行实验分析，因此该思路有进一步探讨的必要，不过显然的是，如果检索结果集的文档之间链接信息较少，则该思路价值不大；设定阈值的做法可以认为是将难点进行了转移，即虽然 K 值无需确定，但是阈值也不易设定。在基础的聚类算法研究中，基于聚类有效性函数的解决方法的算法思想简单，但是需要付出较大的时间开销，不适合检索结果聚类，但是基于遗传算法等优化算法的 K 值确定思路值得期待[24]。

1.2 基于标签的算法

基于标签算法的主要思想是寻找有效的单一特征（即标签），根据单一特征将片段分配到不同的类别中，典型工作是 Zamir 等提出的 STC 算法[3-4]。STC 算法根据后缀树发现共现短语（反复出现的项序列，并不一定具有语法或语义价值）作为文本聚类和生成类别标签的依据；将包含同一高频词序列的文本划分为一个基类；文档集重叠太多的基类被不断地合并，直到无法合并为止。

合并计算使用二进制相似度算法，即定义基类 B_n 和 B_m，如果 $|B_m \cap B_n|/|B_m| > 0.5$ 并且 $|B_m \cap B_n|/|B_n| > 0.5$ 则二者的相似度为 1，表示可以合并。聚类结果显示得分较高的 K 个类，其中类得分通过 $\text{score}(B) = |B| * f(|P|)$ 计算，其中 $|B|$ 为类 B 中的文档数，$|P|$ 为类 B 的标签短语中有意义项的数量，即短语 P 的长度，f 随 $|P|$ 线性增长，如果短语长度大于 6 则 f 保持不变（公式 5）。

$$f(|P|)\begin{cases}0.5 & \text{if } |P| = 1\\ |P| & \text{if } 1 < |P| \leqslant 5\\ 6 & \text{if } |P| > 5\end{cases} \tag{5}$$

STC 算法提出之后，在研究与应用中发现其存在以下不足：后缀树模型是结合英语提出来的，针对中文难以有效抽取关键短语，且易于生成无意义短语；STC 主要应用于网页片段聚类，导致一些没有包含高频短语的文档即使与查询相关也难以出现在类簇中；构建后缀树时易遗漏较长的高质量短语；后缀树占用内存较大；后缀树基类计分合并方法过于简单等，下面是主要的改进方向。

1.2.1 类计分优化 针对 STC 算法存在对类中重叠文档进行重复计分的不足，D. Crabtree 等提出了 ESTC 算法，即将基类 B 的每个文档得分记为 $s/|B|$，类的总分为其中所有文档得分的均值，以 F 值为评价指标，当阈值大于 0.5 时，ESTC 比 STC 提高 50%[25]。Zhang 等通过结合 TF-IDF 和短语的独立性提出新的常见短语计分方法，用以提升聚类的精确度，相较于 STC 算法，平均准确率提高了 5%[26]。

1.2.2 类合并优化 J. Janruang 等[27]通过公式（6）改进了 STC 的合并运算，发现了更为真实的常用短语标签，聚类的平均准确率比 STC 高 10%。其中 A 和 B 是两个基类，$A(d)$ 是 A 类中的片段，$B(d)$ 是 B 类中的片段，$\{a_0, a_1, \cdots, a_n\}$ 是一系列出现在 A 类中的标签短语。Maslowska 提出的 HSTC 算法通过计算 $|C_i \cap C_j|/|C_j| \geqslant \alpha$，$\alpha \in (0.5; 1]$ 发现了类之间的包含关系，使得最终的聚类导航更具层次性[28]。

$$A\oplus B=\left\{\begin{array}{l}a_0\ \oplus\\ a_1=b_0\\ a_2=b_1\\ \vdots\\ a_n=b_{n-1}\\ \oplus\ b_n\end{array}\right\}\ \text{if}(A_{(d)}\subseteq B_{(d)}\ \text{or}\ B_{(d)}\subseteq A_{(d)})\quad(6)$$

1.2.3 候选标签选择 标签是STC算法的基础与质量保证。Hu X等[19]将题名中的项引入聚类标签的选择过程。骆雄武等[29]在候选标签的选择上利用了一些启发式规则，如只包括实词，不包含停用词以及查询词等。Zhang D[30]等通过比较高频项的上下文，基于项的长度和频率定义项的重要性，结合互信息完成项的选择。

针对类簇中难以获得准确有意义的标签，Osinsk[8]提出了“描述优先”的Lingo算法。该算法以优先发现数量有限且有意义的标签为目标，然后将文档分配至标签对应的类簇。Osinsk整合了高频短语发现算法[30]以及潜在语义标引，保证了覆盖所有输入，同时将聚类平均准确率提高至76%。Zeng Huajun等设计的SRC系统[31]尝试利用真实数据集上的训练结果改善标签的选取，将无监督的检索结果聚类问题转换为有监督的突出短语排序问题。SRC整合了TFIDF、项长度n、簇内相似性、类熵以及项独立性计算标签权重，权重值由训练文档获得，试验表明算法复杂度与文档数呈线性关系，得分最高的10个类覆盖了超过50%的文档。

1.2.4 数据结构优化 PAT-tree能够在线性时间内确定关键词频率，张健沛等[32]将其和STC进行了整合，克服了STC处理中文信息的不足。Wang J等[33]将STC和*N-gram*相结合提出了一种新的后缀树，N的设定使得后缀树过滤了一些较长的项，实验表明：10767个STC短语经过3-*gram*最终保留下来的只有5765个，降低了内存空间的开销，比经典STC更快，但是发现的标签比经典STC要短。Zhang D[30]等提出的SHOC（Semantic Hierarchical Online Clustering）算法，采用了后缀数组（Suffix Array）代替后缀树，降低了内存的开销。

1.2.5 基于语义的优化 STC等基于标签的算法高度依赖文档中的高频关键项，忽略了项间的隐含语义关系。对此，许多学者利用潜在语义标引[8,30]的研究成果和各种外部知识源（如Wikipedia，WordNet），提出了基于语义的检索结果聚类算法。

Bellegarda等[34]首先提出了基于LSI的聚类方法，它将最初的“文档—项”空间转化为较低维的“主题—项”空间，并将主题相似矩阵记录在结果文档中，之后根据K个最大的奇异值，将文档分配到最相似的K个主题中。Ciansalvatore[35]等提出了Dynamic SVD聚类，优化了SVD的K值选择，使得SVD能在实际有效的时间内完成，平均F值提高到92.5%。

S. Banerjee等[36]基于Wikipedia完成片段聚类，发现了候选标签和高频项之间的语义关系。实验结果表明：不同算法使用Wiki准确度均有提升，其中层次聚类提升达2倍以上，基于图的聚类准确度最高可达89.56%。Han X等[37]利用维基的语义知识将相关主题转化为一系列维基百科概念，据此构建用于聚类的语义图。相关的工作还有利用同义词词典、网页目录或者同义词发现方法辅助特征选择[38]，SNAKET[39]运用Dmoz网页目录排序从片段提取缺陷语句等。

1.2.6 一点思考 STC解决了传统聚类算法的诸多不足。第一，聚类算法的时间复杂度。STC算法构建后缀树的时间复杂度与句子数呈线性关系，基类的确定和合并也不超过线性时间。在时间复杂度方面STC与经典的K-means相当，但是STC没有K-means算法中难以解决的K值设定问题。第二，软聚类。以STC为代表的基于标签的聚类算法克服了传统聚类算法中普遍存在的硬聚类问题，成功地实现了软聚类，其中STC算法可以将每个片段平均分到2.6个类中[3-4]。第三，准确率。经典STC算法的聚类准确率较K-means提高了30%，改进后的ESTC和HSTC优于经典STC算法，因此有理由认为STC算法族总体上优于以K-means为代表的经典算法。第四，效率优先。需要注意的是，$O(N)$是理想的时间复杂度，但在研究与实际应用中，检索结果集通常较小，因此即使诸如Lingo算法虽然时间复杂度为较高的$O(N^3)$，不过相对于聚类质量的提高，时间上的略微延迟用户也是可接受的。

2 检索结果聚类的问题与未来

目前，检索结果聚类还存在一些不足，比如检索结果聚类的层次还不够完善、对聚类结果输出缺乏预见性、簇的粒度不均匀、聚类标签和内容的不一致性也会影响系统的有用性。据此，未来的研究可以集中于以下方面。

1）提高聚类结果输出有效性。现有的检索结果聚类研究主要集中于检索算法的改进，对结果输出的表现力关注不够[27]。现有的方案有：为类别提供类簇内容的预览[40]；用多文档自动摘要解决表现力问题[41]；B. Stein等[42]根据长尾理论对搜索引擎返回文档进行区分，对相关性排序低的尾部文档进行聚类，进而结合相关性高的文档完成结果展示。近期的聚类系统可用性研究也提供了新视角，Rivadeneira等[43]研究了聚类交互界面，文献[44]为适应移动需求，针对移动端的特点对聚类数、展示结构等方面进行了改进。Giacomo的研究表明[45]可视化可以为

用户带来良好的导航体验，从而表明可视化在该领域的研究与应用将是一个重要方向。

2）特征选择。首先，片段是特征选择的依据，提高搜索引擎返回的片段质量显然有利于获得更有价值的特征[11]。其次，形成更有说服力的特征描述。提出的改进方法包括使用超链接[9]、命名实体、多特征融合[20]等。第三，诱导词的应用。R. Navigli 等[13]利用查询词的共现生成了一系列诱导词，以此提高聚类准确性。检索结果聚类也可以和排名列表（Ranked List）等结合以提高聚类质量[46]。在经典文档聚类算法与 STC 算法优化的路径上，有一个共性的方向就是特征选择的优化，基于 Wikipedia 以及 WordNet 等外源性知识对于两类算法聚类准确性的改进都有明显的价值。

经典文档聚类算法优化研究中的共现信息、链接信息、高频短语、词汇相关度及其融合应用与 STC 算法优化中的候选标签选择也不约而同地走到了一起，表明选择更富语义的特征对于大部分自然语言处理研究具有共性的价值。

3）对用户行为研究成果的挖掘。用户行为研究的成果能促进检索结果聚类和搜索引擎的结合。S. Koshman 等[47]基于 Vivisimo 为期两周的日志分析表明：绝大多数的查询仅包括两个项；绝大多数的查询会话仅有一次查询并且短于一分钟；几乎半数的用户仅浏览单一的类簇，极少部分用户会展开类簇树；11.1% 的检索会话是多任务的，其中包括丰富的查询主题。Gong Xuemei 等[48]基于 Scatter/Gather 系统对检索结果聚类功能进行了有效性评估。这些结合了用户特征研究的启示是：检索结果聚类研究已经利用了大量的特征，用户行为特征的充分挖掘显然有利于提高系统的可用性；聚类不仅依赖于检索结果返回的片段，加入用户个性特征时，还可以形成个性化的聚类检索[49]。

4）多途径的协同研究。采用单一方法往往难以达到预期的效果[11,21]，S. Maiti 等[10]、张刚等[20]的研究表明，将多种特征以及算法协同应用可以有效地提高检索结果聚类的准确率，因此在该领域的研究中应尽可能从多个不同的视角考虑问题。比如 S. Vadrevu 等[50]对聚类的框架进行了调整，提出了三步聚类过程，分别是离线聚类（Offline Clustering），增量聚类（Incremental Clustering）以及实时聚类（Realtime Clustering）；Fred 和 Jain 的研究表明在非预定义结构的情况下合并多个聚类结果有助于类簇识别[51]；基于图划分的聚类[13-14]和利用 LSI 改进的聚类算法[8,30]在聚类准确性等方面有所提高。所以结合各种方法之长，对其进行有机整合，将是结果聚类研究的又一重要方向。□

参考文献

[1] ZHAI C, COHEN W, LAFFERTY J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval [C] //Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2003: 10-17.

[2] AGRAWAL R, GOLLAPUDI S, HALVERSON A, et al. Diversifying search results [C] //Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York: ACM, 2009: 5-14.

[3] ZAMIR O E. Clustering web documents: a phrase-based method for grouping search engine results [D]. Washington: University of Washington, 1999.

[4] ZAMIR O, ETZIONI O. Web document clustering: a feasibility demonstration [C] //Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1998: 46-54.

[5] XU Rui. Survey of clustering algorithms [J]. Neural Networks. New York: IEEE, 2005, 16 (3): 645-678.

[6] ROKACH L. A survey of clustering algorithms [M] // MAIMON O, ROKACH L. Data mining and knowledge discovery handbook. Springer US, 2010: 269-298.

[7] CARPINETO C, OSIŃSKI S, ROMANO G, et al. A survey of web clustering engines [J]. New York: ACM Computing Surveys (CSUR), 2009, 41 (3): 17.

[8] OSIŃSKI S, STEFANOWSKI J, WEISS D. Lingo: search results clustering algorithm based on singular value decomposition [M] // MIECZYSLAW A, SLAWOMIR W, KRZYSZTOF T. Intelligent information processing and web mining. Berlin: Springer-Verlag, 2004: 359-368.

[9] WANG Y, KITSUREGAWA M. Use link-based clustering to improve web search results [C] //Proceedings of the Second International Conference on. New York: IEEE, 2001: 115-124.

[10] MAITI S, SAMANTA D. Clustering web search results to identify information domain [M] // SABNAM S, KUNAL D, GITOSREE K. Emerging Trends in Computing and Communication. Springer, 2014: 291-303.

[11] WANG Fei, LU Yueming, ZHANG Fangwei, et al. A new method based on fuzzy c-means algorithm for search results clustering [M] // YUAN Yuyu, WU Xu, LU Yueming. Trustworthy computing and services. Berlin: Springer Berlin Heidelberg, 2013: 263-270.

[12] DI MARCO A, NAVIGLI R. Clustering web search results with maximum spanning trees [C] //AI * IA 2011: Artificial Intelligence Around Man and Beyond. Berlin: Springer, 2011: 201-212.

[13] NAVIGLI R, CRISAFULLI G. Inducing word senses to improve web search result clustering [C] // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 116-126.

[14] SHA Y, ZHANG G, JIANG H. Text clustering algorithm based on lexical graph [C] //Fuzzy Systems and Knowledge Discovery, Fourth International Conference on. New York: IEEE, 2007: 277-281.

[15] ZHANG Y, FENG B. A co-occurrence based hierarchical method for clustering web search results [C] // Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. New York: IEEE, 2008: 407-410.

[16] HÚSEK D, POKORNY J, REZANKOVA H, et al. Data clustering: from documents to the web [J]. Web Data Management Practices: Emerging Techniques and Technologies, 2006 (2): 1-33.

[17] WANG Y, KITSUREGAWA M. Evaluating contents-link coupled web page clustering for web search results [C] //Proceedings of the Eleventh International Conference on Information and Knowledge Management. New York: ACM, 2002: 499-506.

[18] MAAREK Y, FAGIN R, BEN-SHAUL I, et al. Ephe-meral document clustering for web applications [J]. IBM Research Report, 2000: 1-26.

[19] HU X, SUN N, ZHANG C, et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge [C] //Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 919-928.

[20] 张刚，刘悦，郭嘉丰，等．一种层次化的检索结果聚类方法 [J]．计算机研究与发展，2008，45 (3)：542-547.

[21] STEVEN S, MARTINE DE C, CHRIS C, et al. Clustering web search results using fuzzy ants [J]. International Journal of Intelligent Systems, 2007, 22, 455-474

[22] 夏斌，徐彬．基于超链接信息的搜索引擎检索结果聚类方法研究 [J]．电脑开发与应用，2007，20 (5)：16-17.

[23] CHENG Jia, ZHOU Jingyu, QIU Shuang. Fine-grained topic detection in news search results [C] //Proceedings of the 27th Annual ACM Symposium on Applied Computing. New York: ACM, 2012: 912-917.

[24] 吴夙慧，成颖，郑彦宁，潘云涛．K-means 算法研究综述 [J]．现代图书情报技术，2012 (5)：28-35.

[25] CRABTREE D, GAO X, ANDREAE P. Improving web clustering by cluster selection [C] //The 2005 IEEE/WIC/ACM International Conference on. New York: IEEE, 2005: 172-178.

[26] ZHANG W, XU B, ZHANG W, et al. ISTC: A new method for clustering search results [J]. Wuhan University Journal of Natural Sciences, 2008, 13 (4): 501-504.

[27] JANRUANG J, KREESURADEJ W. A new web search result clustering based on true common phrase label discovery [C] //Computational Intelligence for Modeling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on. New York: IEEE, 2006: 242-242.

[28] MASLOWSKA I. Phrase-based hierarchical clustering of web search results [M]. Berlin: Springer, 2003: 555-562.

[29] 骆雄武，万小军，杨建武，等．基于后缀树的 Web 检索结果聚类标签生成方法 [J]．中文信息学报，2009，23 (2)：83-88.

[30] ZHANG D, DONG Y. Semantic, hierarchical, online clustering of web search results [M] //Advanced Web technologies and applications. Berlin: Springer, 2004: 69-78.

[31] ZENG Huajun, HE Qicai, CHEN Zheng, et al. Learning to cluster web search results [C] //Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2004: 210-217.

[32] 张健沛，刘洋，杨静，等．搜索引擎结果聚类算法研究 [J]．计算机工程，2004，30 (5)：95-97.

[33] WANG J, MO Y, HUANG B, et al. Web search results clustering based on a novel suffix tree structure [M] //Autonomic and trusted computing. Berlin: Springer, 2008: 540-554.

[34] BELLEGARDA J R, BUTZBERGER J W, CHOW Y L, et al. A novel word clustering algorithm based on latent semantic analysis [C] //Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. New York: IEEE, 1996: 172-175.

[35] MECCA G, RAUNICH S, PAPPALARDO A. A new algorithm for clustering search results [J]. Data & Knowledge Engineering, 2007, 62 (3): 504-522.

[36] BANERJEE S, RAMANATHAN K, GUPTA A. Clustering short texts using wikipedia [C] //Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 787-788.

[37] HAN X, ZHAO J. Topic-driven web search result organization by leveraging Wikipedia semantic knowledge [C] //Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010: 1749-1752.

[38] ROLE F, NADIF M. Beyond cluster labeling: semantic inter-

pretation of clusters' contents using a graph representation [J]. Knowledge-Based Systems, 2014, 56: 141-155.

[39] FERRAGINA P, GULLI A. A personalized search engine based on web-snippet hierarchical clustering [J]. Software: Practice and Experience, 2008, 38 (2): 189-225.

[40] OSDIN R, OUNIS I, WHITE R W. Using hierarchical clustering and summarisation approaches for Web retrieval: glasgow at the TREC 2002 interactive track [C]. TREC, 2002.

[41] HARABAGIU S, LACATUSU F. Topic themes for multi-document summarization [C] //Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2005: 202-209.

[42] STEIN B, GOLLUB T, HOPPE D. Beyond precision@ 10: clustering the long tail of web search results [C] // Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM, 2011: 2141-2144.

[43] RIVADENEIRA W, BEDERSON B B. A study of search result clustering interfaces: comparing textual and zoomable user interfaces [J]. Studies, 2003, 21 (5).

[44] CARPINETO, CLAUDIO, et al. Mobile information retrieval with search results clustering: prototypes and evaluations [J]. Journal of the American Society for Information Science and Technology, 2009, 60 (5): 877-895.

[45] DI GIACOMO E, DIDIMO W, GRILLI L, et al. Graph visualization techniques for web clustering engines [J]. Visualization and Computer Graphics, 2007, 13 (2): 294-304.

[46] LEUSKI A, ALLAN J. Improving interactive retrieval by combining ranked list and clustering [C]. RIAO, 2000: 665-681.

[47] KOSHMAN S, SPINK A, JANSEN B J. Web searching on the vivisimo search engine [J]. Journal of the American Society for Information Science and Technology, 2006, 57 (14): 1875-1887.

[48] GONG Xuemei, KE W, KHARE R. Studying scatter/gather browsing for web search [C] // Proceedings of the American Society for Information Science and Technology. New Jersey: Wiley, 2012, 49 (1): 1-4.

[49] CAI K, BU J, CHEN C. An efficient user-oriented clustering of web search results [C] //Computational Science - ICCS 2005. Springer Berlin Heidelberg, 2005: 806-809.

[50] VADREVU S, TEO C H, RAJAN S, et al. Scalable clustering of news search results [C] //Proceedings of the Fourth ACM International Conference on Web Search and Data mining. New York: ACM, 2011: 675-684.

[51] FRED A L N, JAIN A K. Combining multiple clustering using evidence accumulation [J]. Pattern Analysis and Machine Intelligence, 2005, 27 (6): 835-850.

作者简介：柏晗，硕士生。研究方向：信息检索。

成颖，教授，博士生导师。研究方向：信息检索，信息行为。

柯青，副教授。研究方向：人机交互。

收稿日期：2015-03-26

（上接第129页）

术生命周期来考虑，如何将技术生命周期维度加入技术层面专利组合分析模型中也是我们在未来的研究中将继续解决的问题。□

参考文献

[1] ERNST H. Patent portfolios for strategic R&D planning [J]. Journal of Engineer Technology Manage, 1998, 15 (4): 279-308.

[2] ERNST H. Patent information for strategic technology management [J]. World Patent Information, 2003, 25 (3): 233-242.

[3] BROCKHOFF K K. Indicators of firm patent activities [C]. Technology management: the new international language. IEEE, 1991: 476-481.

[4] ERNST H, SOLL J H. An integrated portfolio approach to support market-oriented R&D planning [J]. International Journal of Technology Management, 2003, 26 (5): 540-560.

[5] 李春燕，石荣．专利组合理论研究 [J]．图书情报工作，2009，53 (4)：65-68.

[6] 许高建，胡学钢，王庆人．文本挖掘中的中文分词算法研究及实现 [J]．计算机技术与发展，2007，17 (12)：122-124.

[7] 化柏林．知识抽取中的停用词处理技术 [J]．现代图书情报技术，2008 (8)：48-51.

[8] 姚清耘，刘功申，李翔．基于向量空间模型的文本聚类算法 [J]．计算机工程，2008，34 (18)：39-41.

[9] 黄东流，张旭，刘娅．多维信息分类方法研究——以政府科技管理决策信息为例 [J]．情报杂志，2013，32 (5)：158-165.

作者简介：张世玉，男，1989年生，硕士生。

王伟，男，1958年生，博士，教授，博士生导师。通讯作者。

于跃，男，1988年生，博士生。

付晓燕，女，1992年生。

谭婉君，女，1990年生，硕士生。

陶成琳，女，1990年生，硕士生。

收稿日期：2015-03-02