

数据库、信号与信息处理

## 基于关键词短语聚类的中文搜索结果聚类

麻雪云<sup>1, 2</sup>, 肖诗斌<sup>1, 2</sup>, 王弘蔚<sup>1, 2</sup>, 施水才<sup>1, 2</sup>

1.北京信息科技大学 中文信息处理研究中心, 北京 100101

2.北京拓尔思信息技术股份有限公司, 北京 100101

收稿日期 2008-6-16 修回日期 2008-10-16 网络版发布日期 2009-11-19 接受日期

**摘要** 目前, 搜索结果聚类方法大多数采用基于文档的方法, 不能生成有意义的聚类标签。为了解决这个问题, 提出一种基于关键词短语聚类的中文搜索结果聚类方法, 该方法将名词短语、相关搜索词作为候选聚类标签, 利用C-Value算法、IDF值筛选标签, 然后使用Chameleon算法将标签聚类, 最后将搜索结果划分到最相关的聚类簇。实验证明, 该方法把关键词短语和相关搜索词作为聚类标签, 有效地提高了标签的描述性, 降低了聚类算法的时间复杂度。

**关键词** [搜索结果聚类](#) [关键词短语抽取](#) [C-Value算法](#) [Chameleon算法](#)

**分类号** [TP391](#)

## Chinese search result clustering based on key noun phrase clustering

MA Xue-yun<sup>1, 2</sup>, XIAO Shi-bin<sup>1, 2</sup>, WANG Hong-wei<sup>1, 2</sup>, SHI Shui-cai<sup>1, 2</sup>

1.Chinese Information Processing Research Center, Beijing Information Science and Technology University, Beijing 100101, China

2.Beijing TRS Information Technology Co.Ltd, Beijing 100101, China

### Abstract

Nowadays, the conventional search result clustering methods employ the document-based approach and can not generate clusters with highly readable names. To solve the problem, based on key noun phrase clustering, this paper proposes a method for Chinese search result clustering. First is to extract key phrases from search results, and use the phrases of correlative search as addition. Second is a new label selecting criterion based on C-Value algorithm and the value of IDF. The third is clustering the labels by Chameleon algorithm. Finally, the search result classification has been performed in terms of the results of label clustering. The experiment shows that using key noun phrases and the phrases of correlative search as clustering labels can improve the description of labels and reduce the computation complexity of clustering algorithm.

**Key words** [search result clustering](#) [key noun phrase extraction](#) [C-Value algorithm](#) [Chameleon algorithm](#)

DOI: 10.3778/j.issn.1002-8331.2009.31.035

通讯作者 麻雪云 [maxueyun@mail.bit.edu.cn](mailto:maxueyun@mail.bit.edu.cn)

### 扩展功能

#### 本文信息

- ▶ [Supporting info](#)
- ▶ [PDF\(746KB\)](#)
- ▶ [\[HTML全文\]\(0KB\)](#)
- ▶ [参考文献](#)

#### 服务与反馈

- ▶ [把本文推荐给朋友](#)
- ▶ [加入我的书架](#)
- ▶ [加入引用管理器](#)
- ▶ [复制索引](#)
- ▶ [Email Alert](#)
- ▶ [文章反馈](#)
- ▶ [浏览反馈信息](#)

#### 相关信息

- ▶ [本刊中 包含“搜索结果聚类” 的相关文章](#)
- ▶ [本文作者相关文章](#)

- [麻雪云](#)
- [肖诗斌](#)
- [王弘蔚](#)
- [施水才](#)