# 科技查新中检索词智能抽取系统的设计与实现

王培霞 余海 陈力 王永吉

王培霞,余海,陈力,王永吉. 科技查新中检索词智能抽取系统的设计与实现[J]. 现代图书情报技术,2016,(11):82-93. (peixia@nfs.iscas.ac.cn)

## 应用背景

1. 解决科技查新领域检索词选择时的主观性强、手工工作量大、不规范、费时费力的问题。
2. 利用科技查新过程检出的实时相关语料作为领域知识的来源, 并对语料组成类型与关键词抽取效果之间的关系进行讨论。
3. 实现检索词抽取过程的自动化、智能化、规范化

## 系统设计

面向科技查新领域的检索词智能抽取系统由基于网络爬虫的文献在线检索、检索词智能抽取两部分组成，如图1和图2所示。

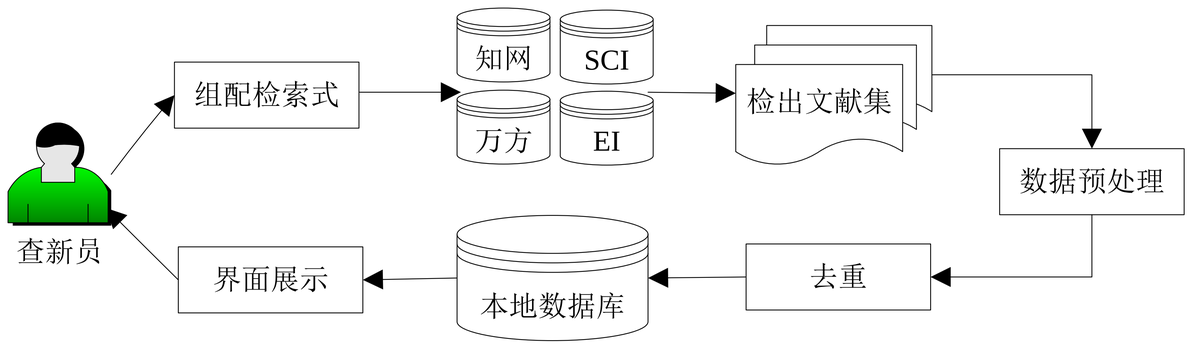


图1 文献在线检索

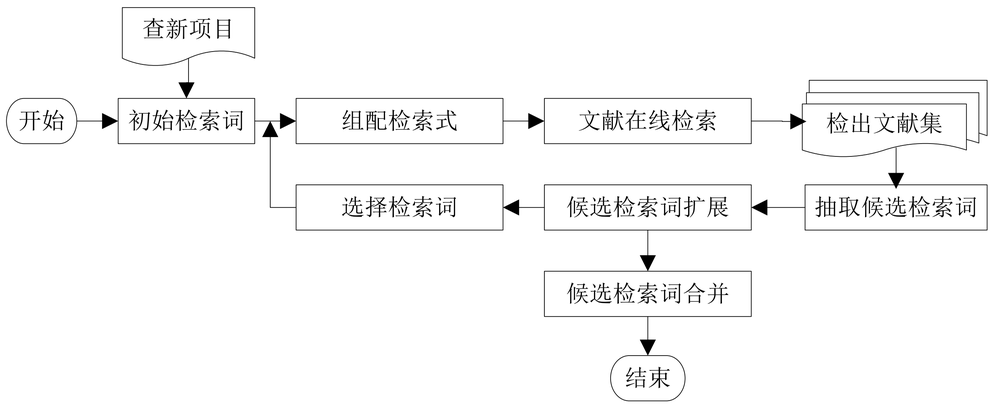


图2 检索词智能抽取

## 方法实现

检索词智能抽取分为：抽取候选检索词和候选检索词的扩展两个步骤。

（1）抽取候选检索词：

$$\mathrm{Score(i)} = (1-\mathrm{d}) + \mathrm{d} \times \sum_{\mathrm{j} \in \mathrm{set(i)}} \frac{\mathrm{tf(i)}}{\sum_{\mathrm{k} \in \mathrm{set(j)}} \mathrm{tf(k)}} \mathrm{score}(\mathrm{t_j})$$

set(i)为词i的共现词，tf(i)为词i的词频

（2）候选检索词扩展：

$$\mathrm{GDC(T)} = \frac{|\mathrm{T}| \log_2 \mathrm{freq(T)} \mathrm{freq(T)}}{\sum_{\mathrm{t} \in \mathrm{T}} \mathrm{freq(t)} \times \mathrm{N}}$$

T 为候选检索词, |T|为其所包含的词语t 的个数

$$\mathrm{DGDC(T)} = \mathrm{tf(T)} \times \mathrm{GDC(T)}$$

## 实验结果及结论

1. 通过与实际查新案例所采用的检索词对比, 发现使用本方法两次迭代后抽取 10 个检索词, 召回率达到 80%。
2. 基于查新过程中检出文献构成的动态相关语料进行检索词的迭代抽取有助于快速、准确锁定绝大部分检索词, 提高检索的效率和效果。