第 37 卷 第 21 期
Vol.37 No.21

计 算 机 工 程
Computer Engineering

2011 年 11 月
November 2011

• 人工智能及识别技术 • 文章编号：1000—3428(2011)21—0155—04 文献标识码：A 中图分类号：TP393

社会媒体网页内容的分割与抽取

解 姝[1]，叶施仁[2]，肖 春[1]

(1. 湘潭大学智能计算与信息处理教育部重点实验室，湖南 湘潭 411105；2. 常州大学信息学院，江苏 常州 213164)

摘 要：为实现社会媒体网页内容的分割与抽取，利用 k-means 算法识别出页面的频繁块并形成一个频繁簇集合，找出该集合中的主题频繁簇，对其中的频繁块结构进行自学习，无需训练样本，即可自动生成抽取规则。实验结果表明，该方法能抽取各种风格的社会媒体网页内容，具有较高的准确率和召回率。

关键词：社会媒体；DOM 结构；k-means 算法；自学习；抽取规则；网页内容抽取

Segmentation and Extraction for Social Media Web Page Content

XIE Shu[1], YE Shi-ren[2], XIAO Chun[1]

(1. Key Laboratory of Intelligent Computing & Information Processing of MOE, Xiangtan University, Xiangtan 411105, China;
2. College of Information, Changzhou University, Changzhou 213164, China)

【Abstract】This paper presents a segmentation and extraction method which does not need any hand-crafted rules and training examples for content-rich pages in social media. It identifies the frequent blocks in page by using k-means algorithm and obtains a collection of frequent clusters. It identifies the topic frequent clusters and induces extraction rules from the frequent blocks in topic frequent clusters through self-supervised approach. Experimental results show that it is efficient and robust for social media Web pages with various styles and layouts with high precision and recall rate.

【Key words】social media; DOM structure; k-means algorithm; self-learning; extraction rule; Web page content extraction

DOI: 10.3969/j.issn.1000-3428.2011.21.053

1 概述

随着互联网的普及，越来越多的互联网用户在博客、论坛、社区等网站上发表用户体验内容，形成了丰富的社会媒体。将社会媒体网页的用户言论抽取出来是进行深度社会媒体分析(如情感分析和主题跟踪)的前提。网页内容抽取规则主要通过 3 种方法构造：(1)人工构造：通过观察网页和网页源码手工构造抽取规则。该方法的抽取准确率高，但需专业人员花费大量时间和人力，只适用于某一类结构相似的网页。(2)有导学习构造：从一系列人工标注的网页中学习抽取规则。该方法不需要专业人员参与，但来源不同的网站内容组织的差异较大，需标注足够多的样本才能训练出有效的抽取模型，开销很大。典型的抽取系统有 Stalker[1]。(3)无导学习构造：利用非人工标注的例子自动获得抽取规则，如通过一系列启发式规则自动识别数据记录边界构造抽取规则[2]。该方法需手工或由其他系统收集一系列相同领域的网页，否则很难归纳出有效的抽取规则。研究表明，无导学习方法的性能尚不理想[3]。

本文提出一种可以识别出社会媒体网页主题频繁簇中的频繁块的网页内容分割与抽取算法，分析归纳了主题频繁簇中所有频繁块的结构特点，并利用树映射生成抽取规则进行信息抽取。该方法主要分为以下 3 个步骤：(1)频繁块的识别；(2)主题频繁簇的识别；(3)抽取规则的生成。

2 社会媒体网页内容分割与抽取系统总体框架

图 1 给出社会媒体网页内容分割与抽取系统的总体框架。首先从网站下载一系列社会媒体页面，进行预处理并解析为对应的 DOM 结构；其次利用 k-means 聚簇方法识别频繁块，得到频繁簇集合，然后对每个频繁簇按照算法进行排序，识别主题频繁簇；最后综合主题频繁簇中频繁块的结构特点归纳学习生成抽取规则，进行信息抽取。其中，频繁块识别模块、主题频繁簇识别模块和抽取规则生成模块是整个系统的核心。

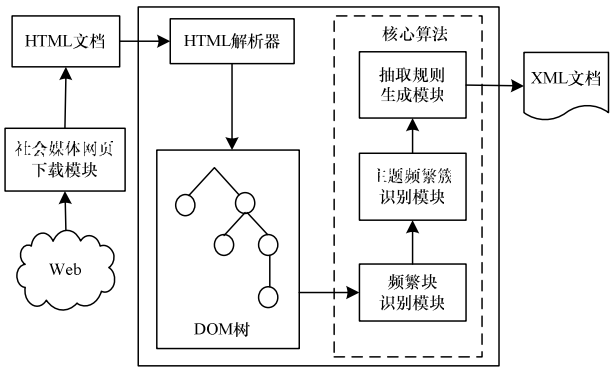

图 1 系统总体框架

3 社会媒体网页内容的分割与抽取

社会媒体页面中不同区域分别表示菜单、导航、版权和内容等部分。这些不同功能的区域所用的 HTML 标记并不完全相同，但局部有些 HTML 标记会反复出现，本文从这些反复出现的部分归纳出稳定的模式来自动抽取所涉及的信息。页面中存在多条相似结构的信息，这样的一条信息称为一个频繁块，频繁块间有时会夹杂一些无关信息，这些信息称为

作者简介：解 姝(1986－)，女，硕士研究生，主研方向：信息抽取；叶施仁，博士；肖 春，副教授、博士

收稿日期：2011-04-21 **E-mail：**xs860123@163.com

非频繁块。每个区域中所有频繁块的集合称为频繁簇，包含待抽取信息的频繁簇称为主题频繁簇。如图 2 所示的部分论坛网页言论区域包含 2 条具有相似结构的帖子，每条帖子为一个频繁块，其中夹杂的广告为非频繁块。由于帖子内容为待抽取内容，因此所有帖子的集合即为主题频繁簇。

图 2 部分论坛网页样例

在网页的 DOM 结构中，同一频繁簇中的频繁块对应的子树有相同的父节点和相似的结构，而与频繁块子树具有相同父节点的非频繁块子树的结构与频繁块的子树结构不同。根据这一特性可识别频繁块。图 3 为对图 2 所示网页进行解析所得的 DOM 结构(部分结构省略)，所有帖子对应由 *Table*、*Tbody*、*Tr*、*Td* 等节点构成的子树集，广告对应一棵由 *P* 等节点构成的子树，它们有相同的父节点 *Table*1。

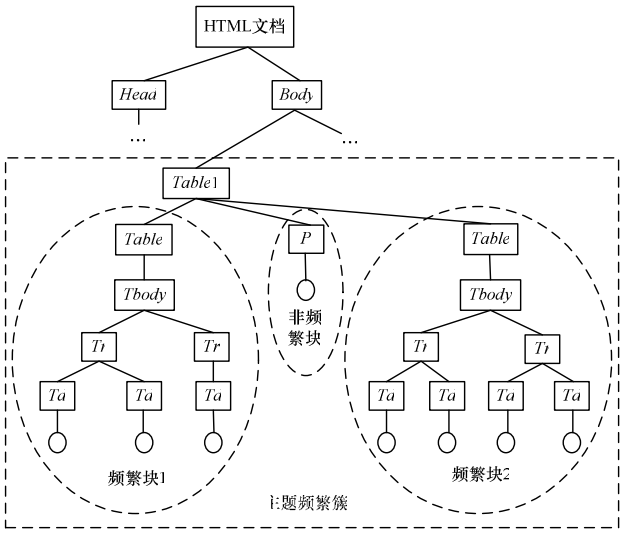

图 3 图 2 网页的 DOM 结构

3.1 频繁块的识别

定义 1 频繁块用来描述菜单、导航、用户言论等内容的重复结构，对应一棵具备以下条件的子树：(1)由动态程序或模板生成，具有稳定的重复结构；(2)如果进一步拆分该子树将影响所表达实体的完整性；(3)浏览器中表现为类似的外观垂直或水平排列。它具有如下特征：(1)频繁块间互为兄弟关系，有共同的父节点，但该父节点下所有的兄弟节点未必都是频繁块；(2)频繁块间具有相似的结构和标记，尤其是它们的主要结构基本相同；(3)频繁块中相应位置的子树表达所涉及对象相同属性的描述。

例如，图 2 中存在 2 个频繁块，分别对应由节点 *Table* 及其子孙节点构成的子树。这 2 棵子树互为兄弟关系，有共同的父节点 *Table*1，且具有相似的结构和标记，都由 *Table*、*Tbody*、*Tr* 和 *Td* 节点构成。虽然一棵子树包含 3 个 *Td* 节点，另一棵子树包含 4 个 *Td* 节点，但它们的主要结构基本相同。可以得出，频繁块间具有很高的相似性，而与非频繁块的相似性很低，本文根据这一特性利用 k-means 聚簇方法识别频繁块。

3.1.1 子树的相似性

本文利用树编辑距离计算子树的相似性。

定义 2 树编辑距离记为 $Td(T_1,T_2)$，其中，T_1、T_2 为树。$Td(T_1,T_2)=\min\{r(p)\}$，其中，p 是将 T_1 转化为 T_2 的一系列树编辑操作。树编辑操作包括：(1)插入一个节点到一棵树中；(2)从一棵树中删除一个节点；(3)将一棵树中的节点转化为其他节点。

计算树编辑距离的过程就是求使 T_1 转化为 T_2 所需树编辑操作的最小次数，将这一求解过程定义为树映射[4]。

令$|T|$为树中节点个数；$T[i]$为树 T 的第 i 个节点，M 为从 T_1 到 T_2 的一个树映射，i 和 j 分别为 T_1 和 T_2 中未连接的节点，$T_1[i]\to T_2[j]$表示 $T_1[i]$转化为 $T_2[j]$，$T_1[i]\to\wedge$表示删除 $T_1[i]$，$\wedge\to T_2[j]$表示插入 $T_2[j]$，则树映射代价定义为：

$$Cost(M)=\sum_{(i,j)\in M}(T_1[i]\to T_2[j])+\sum_{i\in I}r(T_1[i]\to\wedge)+\sum_{j\in I}r(\wedge\to T_2[j]) \tag{1}$$

据此得到最小代价定理：$d(T_1,T_2)=\min\{Cost(M)\}$，其中，$M$ 是从 T_1 到 T_2 的树映射。证明参见文献[5]中定理 1 的证明。

利用最小代价定理计算树编辑距离，将树编辑距离定义为利用树编辑操作实现一棵树到另一棵树所需的最小代价，显然树的相似性与树编辑距离呈反比关系。树编辑距离越小，所需的树编辑操作次数越少，树的相似性越高。

3.1.2 基于 k-means 的频繁块识别

由于频繁块间可能插入非频繁块，因此需要将它们区分开。频繁块与非频繁块具有不同的结构，相似性较低，可采用聚簇方法区分。本文采用 k-means 方法，取 k=2。

对于一个节点 Node，$<ST_1, ST_2,\cdots,ST_m>$为以 Node 的孩子节点为根节点的子树集，$1\leqslant i, j\leqslant m$，$m$ 为 Node 的孩子节点数目，则频繁块的识别过程如下：

步骤 1 初始化聚簇 C_1 和 C_2。计算每个子树 ST_i 的平均树编辑距离 $\overline{Td}(ST_i)$，值最大的子树归为 C_2，其余子树归为 C_1，完成子树的初次分配。$\overline{Td}(ST_i)$ 的计算公式为：

$$\overline{Td}(ST_i)=\frac{\sum_{i\neq j}Td(ST_i,ST_j)}{m-1} \tag{2}$$

其中，$Td(ST_i,ST_j)$根据式(1)和最小代价定理计算。

步骤 2 计算每棵子树 ST_i 相对于 C_1 和 C_2 的平均树编辑距离 $\overline{Td}(ST_i,C_1)$ 和 $\overline{Td}(ST_i,C_2)$。若 ST_i 属于 C_1，则 $\overline{Td}(ST_i,C_1)$ 和 $\overline{Td}(ST_i,C_2)$ 的计算公式为：

$$\overline{Td}(ST_i,C_1)=\frac{\sum_{ST_i,ST_j\in C_1\&ST_i\neq ST_j}Td(ST_i,ST_j)}{\|C_1\|-1} \tag{3}$$

$$\overline{Td}(ST_i,C_2)=\frac{\sum_{ST_i\in C_1\&ST_j\in C_2}Td(ST_i,ST_j)}{\|C_2\|} \tag{4}$$

其中，$\|C_1\|$和$\|C_2\|$表示 C_1 和 C_2 中子树的数目。若 ST_i 属于 C_2，同理计算 $\overline{Td}(ST_i,C_1)$ 和 $\overline{Td}(ST_i,C_2)$。

步骤 3 对每棵子树 ST_i，如果 $\overline{Td}(ST_i,C_1)$ 小于 $\overline{Td}(ST_i,C_2)$，

把 ST_i 归为 C_1；反之，把 ST_i 归为 C_2，完成所有子树的一次重新分配。

步骤 4 重复步骤 2、步骤 3，直至没有子树重新分配。

步骤 5 子树数目大的聚簇中的子树即为所要识别的频繁块。

3.2 主题频繁簇的识别

定义 3 频繁簇对应一个满足以下条件的子树集：(1)频繁块数目大于 1；(2)所有频繁块具有相同的父节点。在图 4 中，每个节点及其子孙节点构成一棵子树，灰色圆圈为频繁块，满足上述条件的子树集有 2 个，得到 2 个频繁簇。

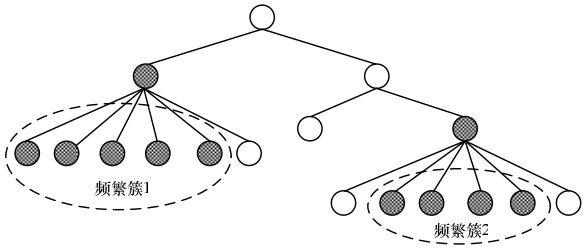

图 4　频繁簇样例

社会媒体网页中通常有多个频繁簇，但待抽取信息仅存在于一个频繁簇中，称为主题频繁簇。其他频繁簇称为非主题频繁簇。通过对多个网页的分析，发现主题频繁簇有以下特点：

(1)频繁块数目较多。一个页面通常包含多个帖子。

(2)平均树编辑距离较大。由于帖子内容占网页内容的比例很大，相对其他非频繁簇中的频繁块，帖子对应的子树结构较大，平均树编辑距离也较大。

(3)平均文本内容较多。由于主题频繁簇中含有帖子内容的句子以及许多描述发帖人信息和发帖时间的描述性句子，因此它们对应 DOM 树中的长文本节点。

非主题频繁簇也可能具备以上特征，但同时具备 3 条特征的情况较少。页面的文本内容主要在主题频繁簇中，非频繁簇一般不具备第(3)个特点。由于在频繁块识别过程中可能将非频繁块误归为频繁块，而非频繁块一般为图片，几乎不包含文本内容，因此在计算每个频繁簇的平均文本内容长度时将产生误差。

为了提高准确率，本文综合考虑上述 3 个特征以及平均文本内容的方差值，得到每个频繁簇的排序值，值最大的即为主题频繁簇。对于频繁簇集合中的任一个频繁簇 $Cluster_i$，$ST(Cluster_i)=<ST_1,ST_2,\cdots,ST_m>$为 $Cluster_i$ 的频繁块子树集，m 为 $Cluster_i$ 中频繁块数目，则主题频繁簇的识别过程如下：

步骤 1 利用 3.1 节的方法计算 $Cluster_i$ 包含的频繁块的个数 m。

步骤 2 计算 $Cluster_i$ 的平均树编辑距离：

$$\overline{Td}(Cluster_i)=\frac{\sum_{ST_j\in ST(Cluster_i)\&ST_k\in ST(Cluster_i)\&ST_j\neq ST_k} Td(ST_j,ST_k)}{m\times(m-1)} \quad (5)$$

步骤 3 计算 $Cluster_i$ 的平均文本内容长度：

$$\overline{ContentLength}(Cluster_i)=\frac{\sum_{ST_j\in ST(Cluster_i)} ContentLength(ST_j)}{m} \quad (6)$$

其中，$ContentLength(ST_j)$表示 ST_j 的文本内容长度。

步骤 4 计算 $Cluster_i$ 的平均文本内容长度的方差值。

$$FangCha(Cluster_i)=\sqrt{\frac{\sum_{ST_j\in ST(Cluster_i)}(ContentLength(ST_j)-\overline{ContentLength}(Cluster_i))^2}{m}} \quad (7)$$

步骤 5 构造目标函数 $Rank(Cluster_i)$以计算每个频繁簇的排序值，值最大的频繁簇即为主题频繁簇。

$$Rank(Cluster_i)=\ln(m\times\overline{Td}(Cluster_i)\times\overline{ContentLength}(Cluster_i)\times\frac{1}{FangCha(Cluster_i)} \quad (8)$$

3.3 抽取规则的生成

主题频繁簇中的频繁块通常具有相同结构，尽管局部存在一些较大的变化，例如一个发帖内容很短，另一个发帖内容通过大量的 HTML 标记来呈现(如包含很多段落和图片标记)，特征如下：(1)相同内容利用稳定的标记，例如，从内容上看，发帖者的财产、发表时间是固定出现在特定位置的。从网页源码上看，利用<class="author…>等修饰独特的频繁块属性。(2)有些网页为节省空间忽略了那些内容为空的属性，或同类型的属性多次出现，导致不同频繁块之间的表达结构出现变化，但这些变化是兼容的，可归纳到相同的框架下。(3)频繁块属性出现的次序是固定的，先后出现在频繁块 A 中的 2 个属性不会在频繁块 B 中颠倒出现，大大限制了待抽取属性的可能组合空间。因此，从频繁块中抽取信息可抽象为待抽取对象通过系列属性来表达。信息抽取任务即为发现这些属性，并将同类型的属性归纳到同一框架下。

如图 2 中帖子内容都包含“发表人昵称”、“发表时间”和“发表内容”等框架且出现在特定位置。根据频繁块的特征，利用 3.1.1 节中的树映射将主题频繁簇中的频繁块中关于相同属性且包含待抽取信息的节点或子树分组，构造抽取规则，最后根据抽取规则进行信息抽取。

抽取规则的生成和信息抽取过程如下：

步骤 1 过滤主题频繁簇中不包含实质内容的节点。

步骤 2 将第 1 个频繁块经过滤后的所有节点的 HTML 属性逐个与其余频繁块经过滤后节点的 HTML 属性进行比较。如果它们的 HTML Tag 相同且节点的深度相同，则将第 1 个频繁块节点的 HTML 属性的出现次数加 1。

步骤 3 对第 1 个频繁块每层的节点进行分析，如果该层每个节点的 HTML 属性出现次数为主题频繁簇中频繁块数目的整数倍，且节点数目最多，将该层节点全部存入列表 L 中。

步骤 4 初始化一个抽取规则集，将每个频繁块的所有节点依次与 L 中的节点比较，如果 2 个节点的 HTML 属性、深度相同，则将该节点加入到抽取规则集中。

步骤 5 利用正则表达式过滤以抽取规则集中的节点为根节点的子树的所有 HTML Tag，得到的文本内容即为待抽取信息的内容。将这些信息以 XML 文件形式输出，抽取信息的片段如下：

```
<!--主题频繁簇-->
  <频繁块>
    <0>央视网友</0>
    <1>2#回复 1# 的帖子因为要买多点几百万一盏的挂灯啊。</1>
    <2>央视网友 发表于 2010-4-22 11:59</2>
    <3>引用回复</3>
  </频繁块>
  <频繁块>
    <0>央视网友</0>
    <1>3#因为办世博可以花钱啊救灾区的钱没有回扣啊!</1>
    <2>央视网友 发表于 2010-4-22 20:36</2>
    <3>引用回复</3>
```

4 实验结果与分析

本文从 10 个网站收集了 500 个论坛网页进行以下 2 个实

验，以验证本文方法的有效性。

(1)主题频繁簇中频繁块识别的实验

该实验从两方面进行评估：1)是否有频繁块遗漏；2)是否有非频繁簇被错分。实验结果如表 1 所示。由表 1 可知，本文方法的平均准确率、平均召回率和平均 F 值都达到了 90%以上，有些甚至达到 100%。原因是初始聚簇时已经将一个子树归为非频繁块簇，即使重新聚簇，非频繁块簇中始终至少包含一棵子树。如果主题频繁簇包含广告等无关内容，则实验结果包含所有的频繁块且不包含非频繁块，准确率、召回率和 F 值都能达到 100%。如果主题频繁簇不包含广告等无关内容，则实验结果不能包含所有的频繁块。

表 1 主题频繁簇中的频繁块识别实验结果 (%)

网页来源	准确率	召回率	F 值
www.xcar.com.cn	94.1	97.2	95.6
www.baa.com.cn	92.5	91.4	91.9
club.autohome.com.cn	100.0	100.0	100.0
bbs.yicars.com	90.8	89.4	90.1
bbs.che168.com	90.3	92.5	91.4
bbs.carcn.net/bbs	100.0	100.0	100.0
bbs.pcauto.com.cn	88.9	87.3	88.1
bbs.360che.com	96.0	93.7	94.8
club.bandao.cn	91.9	90.2	91.0
bbs.greatwallclub.net	92.7	94.3	93.5
平均值	93.7	93.6	93.6

(2)抽取规则生成实验

该实验从两方面进行评估：1)抽取出的信息是否包含帖子所有内容；2)抽取出的信息是否包含其他无关内容。实验结果如表 2 所示。由表 2 可知，本文方法具有较好的抽取性能。即使频繁块具有相似的结构和标记，一些频繁块也可能包含其他频繁块没有的信息。例如，有些用户以注册用户的身份登录发表帖子，有些用户以游客的身份登录发表帖子。注册用户信息一般包括“昵称”、“等级”和“注册日期”等信息，而游客则不包含这些信息。本文方法根据所有频繁块都具有的信息生成抽取规则，所以，抽取的信息中会丢失一些信息，影响了抽取性能。

表 2 抽取规则生成实验结果 (%)

网页来源	准确率	召回率	F 值
www.xcar.com.cn	93.7	97.4	95.5
www.baa.com.cn	100.0	100.0	100.0
club.autohome.com.cn	94.5	93.4	93.9
bbs.yicars.com	92.1	89.9	91.0
bbs.che168.com	89.3	91.2	90.2
bbs.carcn.net/bbs	98.7	95.1	96.9
bbs.pcauto.com.cn	95.9	94.3	95.1
bbs.360che.com	90.0	88.7	89.3
club.bandao.cn	93.9	91.6	92.7
bbs.greatwallclub.net	95.3	92.6	93.9
平均值	94.3	93.4	93.6

5 结束语

本文提出了一种自动抽取社会媒体网页内容的方法。该方法充分利用了 DOM 结构，专门针对网页主题频繁簇的频繁块生成抽取规则，无冗余规则生成，实验结果验证了方法的有效性。本文的进一步工作包括：(1)抽取主题频繁簇的频繁块中的图片信息；(2)将抽取的信息做自然语言处理方面的分析。

参考文献

[1] Muslea I, Minton S, Knoblock C. A Hierarchical Approach to Wrapper Induction[C]//Proceedings of the 3rd International Conference on Autonomous Agents. Seattle, USA: [s. n.], 1999.

[2] 杨 舟, 卓 林. 一种针对商品数据记录的自动抽取方法[J]. 计算机工程, 2010, 36(23): 262-265.

[3] Liu Bing, Grossman R, Zhai Yanhong. Mining Data Record in Web Pages[C]//Proceedings of KDD'03. Washington D. C., USA: [s. n.], 2003: 601-606.

[4] 胡仁龙, 袁春风. 基于重复模式的自动 Web 信息抽取[J]. 计算机工程, 2008, 34(22): 73-76.

[5] Tai Kuochung. The Tree-to-tree Correction Problem[J]. Journal of the Association for Computing Machinery, 1979, 26(3): 422-433.

编辑 张 帆

(上接第 154 页)

同理，R_{21}=0.99；R_{23}=1。因此，EA2 认为 EA 与 EA1 和 EA3 为一个群。EA3 认为与 EA2 为一个群。

当 3 个 Agent 将上述结果上报 MA。MA 依照合群方法，形成群 GP1。GP1 包含 EA1、EA2 和 EA3。其他 Agent 的处理过程与之类似，最终获得的分群结果为：

GP1：EA1，EA2，EA3。

GP2：EA4。

GP3：EA5，EA6，EA7，EA8，EA9。

这里如果不考虑对道路的匹配，将会把 EA1、EA2 分为各自独立的群，同时将 EA3 合并入 GP3 中，这种分群结果在不考虑地形条件下是正确的，但加入地形限制后，这样分群将出现明显的错误。

5 结束语

基于多智能体的战场目标分群方法，考虑战场地形对分群的影响，利用 Agent 的自主性，采用协商的方法，以余弦相似度作为分群的依据，实现对地面目标的分群，实际应用表明本文分群方法是有效的。下一步考虑将推理模块加入到 EA 中，让 EA 根据地形和其自身行动能力限制进行适当预测，以进一步提高分群的准确程度。

参考文献

[1] 黄 雷, 郭 雷. 一种面向态势估计中分群问题的聚类方法[J]. 计算机应用, 2006, 26(5): 1109-1110.

[2] 龙真真, 张 策, 王维平. 基于层次聚类态势估计中的目标分群算法[J]. 弹箭与制导学报, 2009, 29(3): 209-211.

[3] 郭俊文, 覃 征, 贺升平, 等. 机动目标功能的合群算法[J]. 计算机工程, 2006, 32(4): 7-10.

[4] 龙真真, 张 策, 吴伟胜, 等. 基于多帧数据的目标分群算法[J]. 计算机工程, 2009, 35(23): 168-171.

[5] 李伟生. 信息融合中态势估计问题研究[D]. 西安: 西安电子科技大学, 2004.

[6] 李伟生, 王卫星. 实现战术态势估计的一种多 Agent 计划识别方法[J]. 系统工程与电子技术, 2009, 31(3): 613-617.

[7] Rao A S, Murray G. Multi-agent Mental-state Recognition and Its Application to Air-combat Modelling[C]//Proceedings of the 13th Distributed Artificial Intelligence Workshop. Cambridge, USA: [s. n.], 1994: 262-283.

[8] 张 芬, 贾 则, 生佳根, 等. 态势估计中目标分群方法的研究[J]. 电光与控制, 2008, 15(4): 21-23.

编辑 索书志