В. И. Бартош

МЕТОДЫ РАСПОЗНАВАНИЯ ТИПОВ СТАТЕЙ

Рассматривается задача распознавания типов статей. Описываются методы, используемые при распознавании. Проводится сравнительный анализ таких методов.

Введение

Прогресс в области информационных технологий обусловил широкое распространение обработки в реальном времени больших потоков текстовых данных. В современных информационных технологиях роль такой процедуры, как извлечение информации,всё больше возрастает — из-за стремительного увеличения количества неструктурированной информации. Поэтому актуальна проблема создания моделей и алгоритмов, позволяющих эффективно обрабатывать большие потоки данных, особенно в условиях ограниченных временных и других ресурсов. В основе распознавания типов статей лежит классификация текста. В данной статье рассмотрим основные принципы работы двух методов классификации текста.

I. Плоская классификация

Каждый тип статьи, который необходимо установить, назовем измерением. Имея в первом измерении N групп, а во втором M групп, определим (N * M) новых классов, которые представляют собой всевозможные сочетания этих групп. Далее в соответствии с выбранным алгоритмом классификации выполним классификацию текстов. Иллюстрация данного алгоритма для N = 2 и M = 3 представлена на рисунке 1.

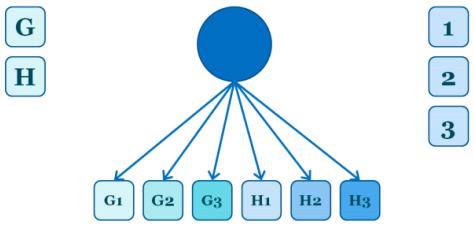

Рис. 1 – Плоская классификация

Этот метод обладает следующими особенностями. Во-первых, при построении модели на каждый класс приходится относительно небольшой объем входных данных. Во-вторых, при больших числах N и M точность классификации может оказаться низкой, так как между некоторыми классами будут существовать лишь незначительные различия. Это затруднит определение класса, к которому относится рассматриваемый объект. Данный метод, на самом деле, является лишь модификацией классической классификации, когда каждому объекту сопоставляется вектор из нескольких значений, обозначающих принадлежность объекта соответствующему классу.

II. Иерархическая классификация

Данный подход заключатся в том, что сначала классификация производится по одному признаку, а затем для получившихся классов независимо друг от друга выполняется классификация по второму признаку. Каждому объекту присваивается несколько классов, в соответствии с количеством извлекаемых признаков. При этом в случае двух измерений возможны два варианта алгоритма, в зависимости от того, по какому признаку классификация производится в первую очередь. Рисунок 2 иллюстрирует данный подход.

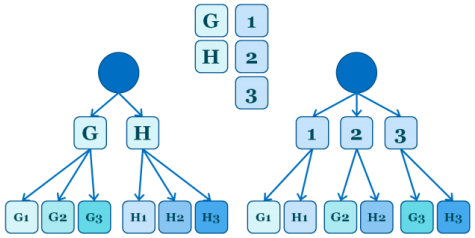

Рис. 2 – Иерархическая классификация

Особенность этого метода заключается в том, что на первом уровне на каждый класс приходится больший объем данных, чем на втором, что может сказаться на точности алгоритма.

III. Выводы

Рассмотренные методы классификации текста позволяют работать со статьями, для определения их типа. Иерархическая классификация обычно сводится к задаче плоской классификации, также у иерархической структуры классов есть возможность налету определять ошибки.

1. Hierarchical Classification of Web Content. / Dumais, S., Chen, H. // Athens, Greece, 2000.

Бартош Владислав Иванович, магистрант кафедры информационных технологий автоматизированных систем БГУИР, vlad.bartosh15@gmail.com.

Научный руководитель: Севернёв Александр Михайлович, доцент кафедры информационных технологий автоматизированных систем, кандидат технических наук, доцент, severnev@bsuir.by.