

# ОПТИМИЗАЦИЯ ГРУППИРОВАНИЯ ОБЪЕКТОВ ПО ЛИНГВИСТИЧЕСКИМ ПРИЗНАКАМ

Н.И. Кекиш, О.С. Карпович-Каспжак

Белорусский государственный университет информатики и радиоэлектроники  
П. Бровки, 6, г. Минск, Беларусь  
телефон: + (375) 29 3128660; e-mail: kekish.n@gmail.com  
web: www.mmts-it.org

Данная статья содержит описание концепций алгоритма группирования объектов по лингвистическим признакам. Рассмотрены варианты применения алгоритма группирования текстов для современных систем публикования текстов. Показано внедрение подхода оптимизации группирования в реальной поисковой информационной системе.

**Ключевые слова –** текст, автоматизация, группирование, интернет, поиск.

## 1 ВВЕДЕНИЕ

Группирование объектов является стандартной в информационных системах операцией. Группирование реализуется с использованием элементарных операции определяющих эквивалентность объектов, как это реализуется в языках программирования и базах данных. Группирование в информационных системах всегда строго формализовано, что достигается использованием характеристик объектов простых для сравнения (длина, размер, порядок и т.д. и их сочетаниями). В случае если группированиес объектов следует производить по параметру не поддающемуся явному сравнению с помощью элементарных операций, то задача сравнения таких объектов становится невозможной с использованием стандартных алгоритмов.

## 2 НЕОДНОЗНАЧНОЕ ГРУППИРОВАНИЕ

Примером задачи группирования с неоднозначным решением является группирование текстов по темам. В ходе решения этой задачи возможные различные варианты решения в зависимости от способа решения и начальных условий. Самые типичные два варианта такие:

1) Имеется фиксированный список групп в которых уже расположены объекты

2) Группы определяются в ходе решения задачи

Анализ двух возможных ситуаций при решении задачи группирования текста с участием человека показывает, что всегда присутствует возможность неоднозначного решения задачи - например, определения одного текста к разным темам в первом варианте, а во втором - составления списка групп самостоятельно во время решения не позволяет решить задачу однозначно в принципе.

Автоматизированное решение задачи разбиения текстов на группы требует наличия формализованного алгоритма. Рассматривая различные пути автоматизированные решения задачи группирования объектов на основе лингвистических признаков, следует проанализировать начальные условия задачи представленные выше. При описании алгоритма необходимо иметь функцию определяющую степень «тематичности» текста в виде числового значения.

## 3 РЕШЕНИЕ ЗАДАЧИ НЕОДНОЗНАЧНОГО ГРУППИРОВАНИЯ ОБЪЕКТОВ

В [1] предлагается формализованный алгоритм сравнения текстов, позволяющий определять степень сходства текстов. Результатом работы алгоритма является коэффициент лежащий в пределах от 0 до 1, определяющий лингвистическую связанность двух текстов. Лингвистическая связанность определяется частотой встречаемых слов в текстах, расчете коэффициента связности текстов используется база синонимов, что улучшает определения коэффициента.

Использования принципов показанных в [3] позволяет построить формализованные алгоритмы для решения задачи группирования объектов на основании лингвистических признаков с начальными условиями упомянутых ранее.

Рассмотрим способы решение задач с начальными условиями:

1) Для определения вхождения в группу необходимо получить коэффициенты сходства рассматриваемого текста с каждым текстом из каждой группы, средний коэффициент сходства будет самый высокий или превышающий определенный порог – будет определять принадлежность текста к группе.

2) Построение матрицы степени сходства текстов «каждый с каждым» с последующим ее обходом позволит получить группы объектов имеющие сходство по лингвистическим признакам.

Для автоматизации именования групп в первом и втором случаях можно использовать часто встречающиеся слова в группе и одновременно редко встречающиеся в других.

Недостатком алгоритмов является присутствия подбираемых экспериментально коэффициентов, определяющих порог, превышение которого определяет выбор группы для объекта, что требует нескольких эксперимен-

гальных запусков с различными значениями для получения качественных результатов.

Достоинством рассмотренных способов группирования объектов на основе их лингвистических признаков является их полная формализация.

#### 4 ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ

Необходимость в структурировании и группировании очевидна. Любой современный сайт, блог, интернет-магазин использует группирования в большей или меньшей степени.

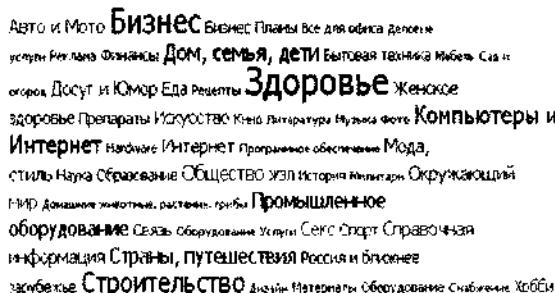


Рис.1. Пример облака тегов для русских текстов

Зачастую группирование в таких системах должно выполняться вручную. Оптимизация группирования страниц интернет сайтов может быть выполнена в виде плагинов, предназначенных для автоматического определения группы (в интернете для этого используется слово «тег») или в разбиении текстов на группы – так называемое построение облака тегов (Рис. 1.).

Другое применение автоматизированного группирования возможно для агрегаторов информации. Современные агрегаторы информации основываются на использовании потоков информации из интернета, целью является поиск по ключевым словам в собранных текстах и получение ссылок на оригиналы текстов. Изначально в таких системах требуется указывать слово для поиска, чтобы получить ссылки на нужные тексты. Использование автоматизированного группирования обеспечит более точ-

ные результаты по темам и предоставит новый функционал обеспечивающий выборку информации по темам (построение облака тегов).

#### 5 РЕАЛИЗАЦИЯ

Представленный подход к программированию автоматизированной группировки объектов был применен в реальной системе поиска по интернет магазинам.

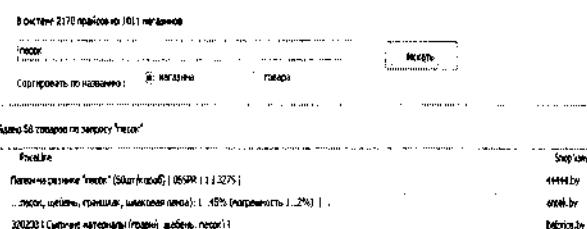


Рис.2. Система поиска по интернет-магазинам.

Главный вид системы поиска показан на (Рис.2.). Используется для определения группы в которую необходимо разместить новый интернет магазин. В качестве текста в реализации поиска по интернет-магазинам используется текст 10 страниц с сайта магазина.

#### ЛИТЕРАТУРА

- [1] Karpowicz-Kasprzak, Olga Оптимизация программы прагматического анализа технических текстов // Известия Белорусской инженерной академии, Минск 2(20)/1 2005.
- [2] Karpowicz-Kasprzak, Olga Precis and review. Comparison of their purposes and indications // VII Международная летняя школа-семинар аспирантов, магистрантов и студентов/ Современные информационные технологии, 2-8 июля, Браслав 2004.
- [3] Дайнек, И. В. Синонимический анализ технического текста / И. В. Дайнек, О.С. Карпович-Каспжак, Н.И. Кекиш Язык и межкультурные коммуникации. - С. 63—65.