УДК 004.9

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ ПОВЫШЕНИЯ ПОЛНОТЫ ПОИСКА ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Зосимов Вячеслав

Миколаївський національний університет ім. В.О. Сухомлинського

zosimovvv@bk.ru

Представлений підхід до підвищення повноти пошуку інформації в мережі Інтернет за рахунок виявлення та групування неунікальних веб-сторінок. Для угруповання веб-сторінок запропоновано використовувати метод кластерного аналізу на основі моделей, побудованих із застосуванням індуктивних алгоритмів.
Ключові слова: Інтерфейс, кластерний аналіз, онтології, індуктивне моделювання, веб-сторінка.

This paper presents an approach to improve the completeness of the information search on the Internet by identifying and grouping of non-unique web pages. For web pages grouping is proposed to use the method of cluster analysis based on models built using inductive algorithms.
Keywords: Interface, cluster analysis, ontology, inductive modeling website.

Представлен подход к повышению полноты поиска информации в сети Интернет за счет выявления и группировки неуникальных веб-страниц. Для группировки веб-страниц предложено использовать метод кластерного анализа на основе моделей, построеных с применением индуктивных алгоритмов.
Ключевые слова: Интерфейс, кластерный анализ, онтологии, индуктивное моделирование, веб-страница.

Введение

Поиск информации в сети Интернет - специфичная задача. Она отличается от поиска информации в тематических каталогах, библиотечных фондах или документах предприятия.

И дело не в том, что в Интернете на порядки больше источников информации и она слабо структурирована.

Основное отличие заключается в том, что большинство информационных ресурсов в сети Интернет в первую очередь используются как инструменты получения прибыли, а информация, размещенная на них - это всего лишь способ привлечения посетителей. Количество посетителей веб-ресурса стало своеобразной валютой или критерием оценки потенциальной прибыли, которую можно извлечь из данного веб ресурса.

В настоящее время один из наиболее популярных способов «заработка в интернете» это ведение персонального или коллективного блога, посвященного определенной тематике. Прибыль авторы блогов получают за счет активности привлеченных на их веб-ресурс пользователей. Это может быть переход по

контекстной рекламе, просмотр рекламных баннеров и видеороликов, переход на сайты продавцов различных товаров через партнерские программы, и т.д.

Для привлечения новых посетителей и удержания уже существующих на веб-ресурсе необходимо регулярное добавление нового качественного и интересного контента. Такой способ развития веб-ресурсов использует лишь небольшой процент авторов. Это обусловлено тем, что на сбор данных и создание качественного текстового, графического или видео-контента уходит много времени. Кроме времени необходимы определенные профессиональные навыки.

Большинство авторов блогов не создают уникальный контент для своих веб-ресурсов. Они получают его путем небольшой переработки уже существующих популярных материалов. Затрачивая при этом минимум времени и усилий.

Если тематика поиска очень востребована, то среди первых десяти результатов выдачи поисковой системы может не оказаться двух принципиально разных материалов. Они все будут в той или иной степени клонами первоисточника. И нет гарантии, что среди этих результатов будет представлен исходный документ. Учитывая то, что обработку исходного документа производят люди не всегда компетентные в теме публикации, качество поиска от этого только ухудшается.

При этом формально с точки зрения критериев полноты и точности поиска, поисковая система показывает отличные результаты. А по сути пользователь лишен альтернативы.

1. Применение кластерного анализа при поиске информации в сети Интернет

В качестве одного из подходов к решению данной проблемы в статье предложен метод кластерного анализа содержимого веб-страниц для выявления и объединения обработанных клонов оригинала в отдельные группы. Пользователь сможет, просмотрев один или несколько документов из определенной группы, решить для себя, была ли полезна и исчерпывающа информация из группы или стоит перейти к просмотру следующей.

Формальная постановка задачи кластеризации.

Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $p(x,x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $\alpha : X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации.

Кластеризация отличается от классификации тем, что метки исходных объектов y_i изначально не заданы, и даже может быть неизвестно само множество Y.[1]

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

- не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты. Следовательно, для определения качества кластеризации требуется эксперт предметной области, который бы мог оценить осмысленность выделения кластеров.
- число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. Это справедливо только для методов дискриминации, так как в методах кластеризации выделение кластеров идёт за счёт формализованного подхода на основе мер близости.

Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом. Но стоит отметить, что есть ряд рекомендаций к выбору мер близости для различных задач. [1]

Существует несколько поисковых систем, в которых применяется кластерный анализ для более комфортного представления результатов поиска пользователю.

Механизмы, использующие кластерный анализ обеспечивают лучшую презентацию результатов поиска, потому что организуют их в структуру. Этот метод заключается в назначении определенных категорий или тематик документам и результатам поиска. Понятие кластер означает множество, совокупность, связку или просто группу. Кластеризация направлена на - насколько это возможно - сортировку результатов поиска в одну или несколько таких групп, и таким образом извлекает группы из всех результатов.

Условием успеха является предварительное определение групп, а также категорий, тематик или слоев, которые, в свою очередь, определяются ключевыми словами и профессиональными понятиями. Для того, чтобы документ мог быть назначен группе, он должен быть правильно классифицирован. Реализация этого намерения не всегда оказывается достаточно простой. Чтобы ее сделать, поисковая система читает, после чего

исследует данные и метаданные документа. Также анализирует содержание документа на основе статистических расчетов (принимает во внимание частоту появления букв, слогов и слов, порядок фраз, а также длины слов и предложений), или использует алгоритмы лингвистического анализа. Чем более точны эти данные, тем точнее можно выделить документ в определенной группе [2].

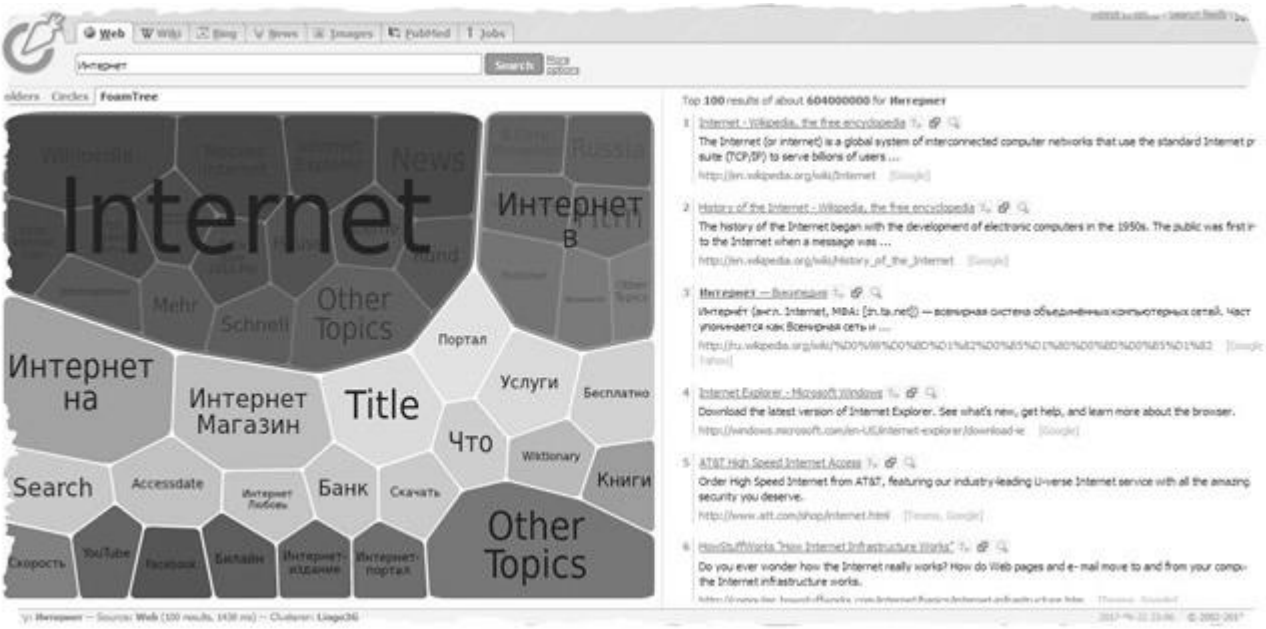

Рис. 1. Страница поисковой выдачи с использованием кластеров.

2. Подход к построению кластеров с применением индуктивных алгоритмов

В рассмотренной в статье задаче нет необходимости заранее определять оптимальное число кластеров. Это число будет разным для каждого поискового запроса и зависеть от количества первоисточников, количества результатов, критериев кластеризации.

Для качественного разбиения веб-сраниц на группы необходимо построить модели кластеризации. В качестве инструмента для построения моделей кластеризации был выбран обобщенный итерационный алгоритм метода группового учета аргументов. Это алгоритм показал высокие результаты при решении задач построения моделей ранжирования для поисковых систем [3].

В данной задаче основной проблемой является выявление характерных признаков, на основе которых будет строиться модель кластеризации. Для различных результатов поиска будет целесообразно применять отдельные признаки.

Возможным решением может быть создание онтологий с описанием различных сфер жизнедеятельности, к которым может относиться введенный

пользователь запрос. Каждая онтология будет содержать определенный набор признаков кластеризации и модели кластеризации [4].

Для повышения эффективности необходимо добавить механизмы обучения для пополнения о обновления признаков кластеризации.

Метод группового учета аргументов основан на принципах теории обучения и самоорганизации, в частности, на принципе массовой «селекции» или самоорганизующемся направленном переборе всевозможных вариантов построения решающего правила классификации. Задача построения решающего правила в МГУА представляется как задача индуктивного построения модели, усложняющейся в процессе работы алгоритма.

В работе рассматривается класс задач моделирования, который содержит информацию про n измерений m входных переменных (признаков) $X[n \times m]$ и одной выходной переменной $y[n \times 1]$. Необходимо найти модель зависимости вход-выход.

Искомая с помощью МГУА модель для рассматриваемой в данной работе задачи будет представлена в классе подмножеств одночленов полинома Колмогорова-Габбора:

$$y(x_1,\ldots,x_m) = \theta_0 + \sum_{i=1}^{m}\theta_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m}\theta_{ij}x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\theta_{ijk}x_i x_j x_k + \ldots , \quad (1)$$

где $\theta = \{\theta_0,\ \theta_i,\ \theta_{ij},\ \theta_{ijk},\ldots\}$ – вектор коэффициентов.

Для определения вектора коэффициентов математической модели по выборке данных применяется метод наименьших квадратов (МНК).

Задача построения модели с выбором ее структуры и оценки параметров сводится к формированию по выборке экспериментальных данных некоторого множества Ф моделей-кандидатов $f \in \Phi$ различной структуры в классе линейной по параметрам функции (2):

$$\hat{y}_f = f(X, \hat{\theta}_f) \quad (2)$$

и поиска оптимальной модели из этого множества Ф как решение задачи дискретной оптимизации при условии минимума внешнего критерия селекции $CR(\cdot)$:

$$f^* = \arg\min_{f \in \Phi} CR(y,\ f(X, \hat{\theta}_f)), \quad (3)$$

В роли критерия селекции будем использовать критерий регулярности, который основан на разбиении выборки на обучающую (А) и проверочную (В):

$$AR_{B|A} = \left\| y_B - \hat{y}_{B|A} \right\|^2 = \left\| y_B - X_B \hat{\theta}_A \right\|^2 \quad (4)$$

В работе будет рассматриваться обобщенный итерационный алгоритм (ОИА) – гибридный алгоритм МГУА, который объединяет свойства комбинаторного и многорядного алгоритмов, исключая при этом их недостатки, перечисленные выше [5].

3. Выводы

Группировка неуникальных веб-страниц, созданных путем переработки оригинального текста позволит повысить полноту поиска информации в сети Интернет за счет уменьшения количества схожих материалов, а значит, увеличения количества альтернатив на первых страницах поисковой выдачи.

Литература

1. http://wikipedia.org.

2. http://search.carrotsearch.com/

3. Zosimov V., Stepashko V., Bulgakova O. Inductive building of search results ranking models to enhance the relevance of the text information retrieval. – Proc. of the 26th Intern. Workshop “Database and Expert Systems Applications, 1-4 Sept., Valencia, Spain / Ed. by Markus Spies at al. – Los Alamitos: IEEE Computer Society, 2015. – 316 p. / – P. 291-295. – ISSN: 1529-4188.

4. Antoniou, G., Franconia, E., & van Harmelen, F. (2005). Introduction to Semantic Web Ontology

5. Stepashko V.S., Bulgakova O.S. Generalized iterative algorithm of the group method of data handling // USiM. – 2013. – № 2. – P: 5–18.