УДК 004.9

Головащенко Р.А.

студент магистратуры

2 курс, факультет «Информатика и системы управления»

Московский государственный технический университет им. Н. Э. Баумана

Россия, г. Москва

Степанов А.В.

студент магистратуры

2 курс, факультет «Информатика и системы управления»

Московский государственный технический университет им. Н. Э. Баумана

Россия, г. Москва

Ларин А.А.

студент магистратуры

2 курс, факультет «Информатика и системы управления»

Московский государственный технический университет им. Н. Э. Баумана

Россия, г. Москва

СЕМАНТИЧЕСКАЯ ЗАВИСИМОСТЬ ДЛЯ ВЫЯВЛЕНИЯ ЗАКОНОМЕРНОСТЕЙ НОВОСТНЫХ КЛАСТЕРОВ

Аннотация

Статья посвящена исследованию семантическому анализу для выявления закономерностей кластеров. В современном мире информация представляет собой ресурс, обладание которым приносит ценную информацию для реализаций проектов, стартапор или компаний. Данные инструменты позволяют структурировать важные стадии для реализация методов принятий решений. К кластерам относят ключевые слова или значения из семантической сети, которые имеют свою подструктуру

ключевых значений. Закономерности данных новостных событиях определяются зависимостью количества просмотров от количества кластера.

Ключевые слова: кластер, новость, динамика, пуассоновское распределение, экспоненциальная закономерность, ключевые значения.

Golovashenko R.A.
Master, 2 year
Department of Information Processing and Management Systems
Bauman Moscow State Technical University
Russia, Moscow
Stepanov A.V.
Master, 2 year
Department of Information Processing and Management Systems
Bauman Moscow State Technical University
Russia, Moscow
Larin A.A.
Master, 2 year
Department of Information Processing and Management Systems
Bauman Moscow State Technical University
Russia, Moscow

SEMANTIC DEPENDENCE FOR DETECTION THE REGULARITIES OF NEWS CLUSTERS

Annotation. The article is devoted to the investigation of semantic analysis for revealing the regularities of clusters. In the modern world, information is a resource, the possession of which brings valuable information for the implementation of projects, start-ups or companies. These tools allow you to structure important stages for the implementation of decision-making methods. Clusters include keywords or values from a semantic network that have their own

key substructure. The patterns of these news events are determined by the dependence of the number of views on the number of clusters.
Key words: cluster, news, dynamics, Poisson distribution, exponential growth, keys value

В работе [1] описан метод подхода основанный на кластеризации. Принцип метода таков, что каждое новостное сообщение относится к одному из k – кластеров. Так как, авторы применяют разработанный метод для анализа в социальных сетей, в моделях мониторинга новостных сообщений, учитывается структурная составляющая. Кластеры представлены в виде векторов тематик, а именно ключевых слов в новостном сообщении.

При поступлении нового новостного сообщения, ключевые слова сравниваются с каждым кластером, которые задаются с помощью построения семантической сетью [2]. На вход подается текстовая информация, которая обрабатывается лингвистически, а именно, определяется множество лексем, множество векторов морфологической информации, множество векторов статической информации о лексемах, множество графов семантической окрестности термина. В данном случае кластер из семантической сети, является концептуальным объектом знака «величина».

Для расчета кластера C_j рассчитывается содержательная близость новостного сообщения $Sim(S_i, C_j)$ посредством сравнения векторов терминов и кластера с использованием косинусной меры.

В конечном итоге кластер C^* рассчитывается следующим образом:

$$C^* = \sum_i \left(\frac{n_i}{n}\right) \arg\max_{C_j} (Sim(S_i, C_j)) \quad (1)$$

где n_i число сообщений в классе i; n – число сообщений в контексте; C_j – кластеры; $Sim(S_i, C_j)$ – содержательная близость новостного сообщения.

По обнаружению популярности новостных данных, была разработана математическая модель [3]. Выявление динамики популярности

осуществляется за счет кликов новостных данных, за определенное количество времени. Динамика выявляется по Пуассоновскому распределению, степенное количество просмотров за среднее статическое время t.

Мы рассматриваем кластеры, как ключевые слова и каждый кластер имеет свой класс ключевых слов. Для новостного ресурса РБК, существует тысяча кластеров и больше ста тысяч показов по поисковой системе Яндекс. После результата кластеризации получается график, зависимости количества просмотров от кластеров, описывающий экспоненциальную закономерность, а именно экспоненциальное падение стремящаяся к бесконечности с допустимым значением.

Чем больше кластеров в новостном потоке, тем меньше просмотров; чем меньше кластеров в новостном потоке, тем больше просмотров (см. рис. 1.). Это говорится о том, что человек не способен физически одновременно анализировать все тысячи кластеров, да и к тому же, для каждой предметной области требуется несколько десятков кластеров. Итоговый алгоритм поведения новостных данных с кластеризацией:

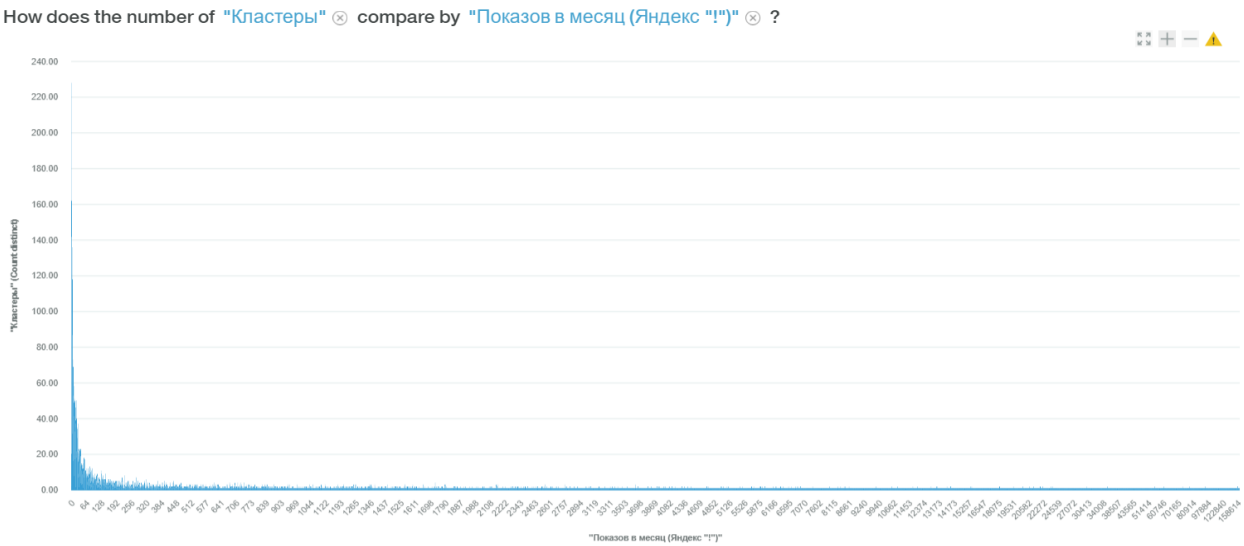

Рис. 1. Графическая зависимость кластеров от показов в месяц

$$C^* = |e^{-P_{ij}}| \quad (2)$$

Экспоненциальный рост или падение, описывает при следующих условиях:

$$F_i = \begin{Bmatrix} \sum C^* = 1000 \\ P_i < 64 \end{Bmatrix} \quad (3)$$

$$F_j = \begin{Bmatrix} \sum C^* < 1000 \\ P_j > 64 \end{Bmatrix} \quad (4)$$

где F_i –экспоненциальная зависимость кластера от количества показов в месяц I - контекста; C^* – количество кластеров; P_i – количество показов в контексте i.

Использованные источники

1. Aggarwal C. C., Subbian K. Event detection in social streams //Proceedings of the 2012 SIAM international conference on data mining. – Society for Industrial and Applied Mathematics, 2012. – С. 624-635.
2. Найханова Л. В., Аюшеева Н. Н., Хаптахаева Н. Б. Построение семантической сети предметной области на основе извлечения знаний из научного текста //Известия высших учебных заведений. Поволжский регион. Технические науки. – 2007. – №. 4.
3. Ratkiewicz J. et al. Characterizing and modeling the dynamics of online popularity //Physical review letters. – 2010. – Т. 105. – №. 15. – С. 158701.