Литература

1. Салин В.С., Папшев С.В., Сытник А.А. Практическое применение метода BorderFlow в задаче автоматизированной семантической кластеризации веб-сайта. // Научно-методический журнал «Информатизация образования и науки» № 3(27)/2015. ФГАУ ГНИИ ИТТ «Информика». С. 65-73.
2. Федеральное агентство по техническому регулированию и метрологии ГОСТ-28806-90: Качество программных средств. Термины и определения // Информационный портал по стандартизации. – Стандартинформ, 2017 – Режим доступа: http://standard.gost.ru/ (дата обращения: 10.01.2017).
3. Alexander A. Sytnik, Sergey V. Papshev. Semantic Segmentation of Hypertext on the Basis of Automata Model. International Journal of Computing Anticipatory Systems, v. 28, 2014, D.M. Dubois (Ed.), CHAOS, Liège, Belgium, ISSN 1373-5411, ISBN 2-930396-17-2. P.109-115.
4. Сытник А.А., Шульга Т.Э. Математические модели адаптивных дискретных систем. Монография //Саратов: Сарат. гос. техн. ун-т, 2015. 272с. ISBN 978-5-433-2947-2.
5. Сытник А.А. Перечислимость при восстановлении поведения автоматов //Доклады РАН. 1993. Т.238. N1. С.25-26
6. Богомолов А.М., Твердохлебов В.А. Диагностика сложных систем. Киев. Наукова Думка. 1974. 128 с.
7. Богомолов А.М., Твердохлебов В.А. Целенаправленное поведение автоматов. Киев. Наукова Думка. 1975. 123 с.
8. Сытник А.А. Методы и модели восстановления поведения автоматов. //Автоматика и телемеханика. 1992. N 11.

А.А. Сытник, С.В. Папшев, Т.Э. Шульга

ОБ ОДНОМ ПОХОДЕ К СЕМАНТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

(Саратовский государственный технический университет имени Гагарина Ю.А)

Аннотация – В статье предлагается решение актуальной проблемы семантической кластеризации нетекстовых веб-документов за счет использования статистики посещения и гиперссылок. Результаты дополняют известные методы семантической кластеризации текстовых документов и предоставить возможность классифицировать текстовые и нетекстовые объекты в рамках единой системы на предварительном этапе интеллектуальной обработки данных на основе автоматных моделей.

Ключевые слова – Семантическая кластеризация, семантический веб, гипертекст, нетекстовый документ, автомат, система, модель; алгоритм.

Abstract: The article proposes the solution of the actual problem of semantic clustering of non-text web documents by using statistics of visits and hyperlinks. The results complement the known methods of semantic clustering of text documents and provide an opportunity to classify text and non-text objects within a single system at

the preliminary stage of intellectual data processing on the basis finite-automaton models

Keyword: Semantic clustering, semantic web, hypertext, non-text document, automata, system, model, algorithm.

Объемы информации, представленной в сети Интернет, постоянно растут, экспоненциально увеличивается количество веб-сайтов в сети. Обработка больших объемов информации с целью эффективного извлечения требующихся данных предполагает использование специализированных программных средств поиска и интеллектуального анализа данных (Н. Шедболт, В. Холл, Т. Бернерс-Ли). На предварительных этапах, многие исследователи применяют подход с группированием веб-документов по тематике или близости по смыслу, который принято называть **семантической кластеризацией**.

Решение проблем семантической кластеризации рассматривалось многими зарубежными и отечественными исследователями (О. Замир, О. Эционi, Дж.О. Педерсен, Б. Штайн, Д. Вайсс, Дж. Стефановски, С. Озински, И. Масловска, Б.В. Крофт, М.А. Хёрст, П. Феррагина, С. Карпинето, Д.В. Михайлов, Г.М. Емельянов и другие). В работах указанных авторов, задача семантической кластеризации решается путем смысловой обработки текстов, извлеченных из соответствующих веб-документов. Вместе с тем, вопрос кластеризации нетекстовых веб-документов (например, графических) остается малоизученным, оставаясь при этом актуальным в анализе веб-сайтов. В таких случаях, известные текстовые методы малоэффективны, и для определения семантической близости между нетекстовыми веб-документами требуются новые подходы.

В этой связи, как представляется, может оказаться полезным подход, связанный с учётом статистики числа обращений пользователей к нетекстовым документам. В частности, подобный учёт может расширить возможности графовых моделей в решении задачи семантической кластеризации.

В данной работе предлагается решение проблемы семантической кластеризации нетекстовых веб-документов за счет использования статистики посещения и гиперссылок. Результаты исследования дополняют известные методы семантической кластеризации текстовых документов и предоставить возможность классифицировать текстовые и нетекстовые объекты в рамках единой системы на предварительном этапе интеллектуальной обработки данных. Изложенное позволяет сформулировать следующую цель работы.

В результате проведенного исследования разработана математическая модель гипертекстовой структуры в виде взвешенного редуцированного графа и предложен метод семантической кластеризации гипертекстовой структуры, использующий данные статистики переходов пользователей между страницами сайта.

Метод не анализирует текстовое содержание документов, не требует полнотекстовой индексации и последующего поиска по индексу, что отличает от традиционных методов текстовой кластеризации и повышает его эффективность. Кроме того, предлагаемый метод позволяет осуществить семантическую

привязку нетекстовых документов, что не позволяют делать традиционные методы.

Разработана программная система для автоматизированного построения модели гипертекстовой структуры в виде взвешенного редуцированного графа для гипертекстовой структуры. Данная система может применяться как инструмент для решения прикладных задач в областях проектирования, разработки веб-сайтов, интеллектуальном анализе данных в задачах из области семантического веба.

Кратко перечислим основные полученные результаты:

1) Разработана математическая модель гипертекстовой структуры в виде взвешенного редуцированного графа, отличающаяся учётом статистики наиболее частых переходов пользователей между узлами гипертекста за заданные промежутки времени.
2) Разработан алгоритм вычисления весов дуг графа по данным статистики переходов пользователей между веб-страницами за заданные промежутки времени. Алгоритм применяется в предложенной модели при построении взвешенного графа веб-сайта.
3) Предложена методика оценки эффективности нового метода семантической кластеризации соотнесения найденных кластеров нетекстовых веб-документов с кластерами текстовых документов. Методика включает расчет численных показателей, а также оценки соответствия нетекстовых документов известной тематике.
4) Разработан комплекс программ для автоматизированного построения математической модели веб-сайта с учётом статистики обращений к веб-документам и её последующей кластеризации. Комплекс программ также позволяет рассчитывать численные показатели эффективности нетекстовой кластеризации при сопоставлении с кластерами текстовых документов.
5) На основании предложенной математической модели и разработанного комплекса программ, показана эффективность кластеризации нетекстовых документов с учётом статистики переходов на трёх тестовых примерах реальных веб-сайтов с различной долей нетекстовых веб-документов в них.

На основании разработанных моделей, методов и комплекса программ, предложены практические рекомендации по использованию результатов исследования в решении прикладных задач: в области семантического веба, результаты могут использоваться для программного семантического анализа кластеров страниц веб-документов, в области веб-разработки и проектировании результаты исследования могут содействовать решению задач построения адаптивной навигации, реинжиниринга веб-сайта и оптимизации его логической структуры.

Литература

1. Салин В.С., Папшев С.В., Сытник А.А. Практическое применение метода BorderFlow в задаче автоматизированной семантической кластеризации

веб-сайта. // Научно-методический журнал «Информатизация образования и науки» № 3(27)/2015. ФГАУ ГНИИ ИТТ «Информика». С. 65-73.

2. Федеральное агентство по техническому регулированию и метрологии ГОСТ-28806-90: Качество программных средств. Термины и определения // Информационный портал по стандартизации. – Стандартинформ, 2017 – Режим доступа: http://standard.gost.ru/ (дата обращения: 10.01.2017).

3. K. Sridevi, R. Umarani, V.Selvi. An Analysis of Web Document Clustering Algorithms. International Journal of Science and Technology. Volume 1 No.6, December 2011, pp.: 275 – 282.

4. Ngomo, A.C.N., Lyko, K., Christen, V.: Coala-correlation-aware active learning of link specifications. In: The Semantic Web: Semantics and Big Data, pp. 442–456.Springer (2013).

5. Alexander A. Sytnik, Sergey V. Papshev. Semantic Segmentation of Hypertext on the Basis of Automata Model. International Journal of Computing Anticipatory Systems, v. 28, 2014, D.M. Dubois (Ed.), CHAOS, Liège, Belgium, ISSN 1373-5411, ISBN 2-930396-17-2. P.109-115.

6. Сытник А.А., Шульга Т.Э. Математические модели адаптивных дискретных систем. Монография //Саратов: Сарат. гос. техн. ун-т, 2015. 272с. ISBN 978-5-433-2947-2.

7. Сытник А.А. Перечислимость при восстановлении поведения автоматов //Доклады РАН. 1993. Т.238. N1. С.25-26

А.А. Санталов, Д.А. Жуков

ДИАГНОСТИКА ТЕХНИЧЕСКОГО СОСТОЯНИЯ СИСТЕМЫ С ПРИМЕНЕНИЕМ НЕЙРОСЕТЕВЫХ МЕТОДОВ

(Ульяновский государственный технический университет)

Исследовалась эффективность применения нейросетей при диагностике технического состояния системы на примере станции водоочистки. В качестве исходных данных использовались семь показателей функционирования системы (физико-химические параметры водоисточника и дозы реагентов для очистки) и состояние системы: система исправна, если показатели качества очищенной (питьевой) воды лежат в допустимых пределах. Задача состоит в разработке нейронной сети, предсказывающей состояние системы по семи заданным показателям функционирования, подбором параметров сети и оценкой эффективности ее прогнозов с использованием метода кросс-валидации.

Сформулированная задача является задачей бинарной классификации, и для ее решения удобно использовать нейронную сеть, настраиваемую функцией patternnet из пакета MATLAB [1]. В качестве алгоритма обучения нейронной сети был использован метод сопряженных градиентов, отличающийся о высокой сходимостью и малыми затратами памяти [2,3].