УДК 004.89

СБОР ГЕОЗАВИСИМЫХ ДАННЫХ ИЗ НОВОСТНЫХ РЕСУРСОВ

Авдюшина А. Е. (Университет ИТМО),
Научный руководитель – доктор технических наук, профессор Бессмертный И.А.
(Университет ИТМО)

Введение. Активное развитие информационных технологий распространено на различны виды информации, в том числе на географическое положение объектов на планете. Чаще всего для хранения таких структурированных данных используются географические информационные системы. В современном информационном пространстве распространено большое количество неструктурированных данных, которые также имеют географические составляющие, для которых необходимо применять поиск и индексацию [1]. Известны в литературе исследования географического информационного поиска [1], которые используют тематическое моделирование и метаданные изображений в социальных сетях для формирования контекста территории, а также исследования [2], которые рассматривают основные проблемы информационного поиска на примере новостных ресурсов. Подобные данные могут быть использованы в совокупности для получения более полной и устойчивой модели поиска, решения актуальной задачи представленной в этой работе.

Основная часть. Область или система поиска географической информации (GIR) – это широкое понятие, которое подразумевает извлечение географической информации из различных наборов данных, таким как электронная коммерция, корпоративные открытые документы, новости, социальные сети и т.д. [4]. Системы GIR представляют собой поисковые инструменты, которые работают не с обычными текстовыми запросами, а с запросами привязки к местоположению, методы геокодирования - определения имен местоположений и связанных с ними координат. Как любая поисковая система GIR должна проводить предварительную индексацию структурированных, но чаще неструктурированных донных, совмещаю текстовую и географическую информацию. Поиск включает процесс индексации как предобработку больших объемов информации. Индексация работает с индексируемым элементом методами простого или выводного извлечения, а также с применением интеллектуального анализа. В поисковых системах GIR чаще всего сочетается дедуктивное индексирование и интеллектуальное индексирование. Например, дедуктивное индексирование извлекает географическую информацию из текста, а интеллектуальное индексирование – из видео или фото объектов анализа.

Методы информационного поиска все чаще используются в междисциплинарном контексте, потому что они хорошо себя зарекомендовали в определении содержания документов. Интересным направлением GIR является определение географической информации из новостных статей, связанных с конкретным местоположением.

С другой стороны, тематическое моделирование как метод машинного обучения используется для извлечения тем из больших набора различных текстовых документов. Двумя наиболее часто используемыми методами тематического моделирования или их усовершенствованиями являются скрытое распределение Дирихле (LDA) и вероятностный латентный семантический анализ (PLSA).

Следовательно, актуально разработать подход для объединения данных новостных порталов в кластера, позволяющие извлекать список тем. В данном работе для получения релевантного ответа при поиске событий по определенному местоположения предлагается использоваться методы GIR и пространственной кластеризации, извлекать именованные сущности.

В этой работе исследуется комплексное использование тематического моделирования и GIR для разработки алгоритма тематик в различных районах Санкт-Петербурга. Алгоритм обработки наборов данных новостей по поиску в них географической информации и тематическому моделированию следующий:

1. собрать и предварительно обработать набор новостных данных для удаления стоп-слов, знаков препинания и специальных символов, также провести нормализацию текста методами лемматизации;

2. определить для каждой новостной статьи ее местоположение с применяем методов распознавания именованных объектов, геокодирования и картографических сервисов;

4. провести тематического моделирования в наборе данных новостных порталов с применением метода машинного обучения LDA;

5. полученные темы новостных данных сохранить для дальнейшего использования, провести их анализ, например, определить наиболее распространенные темы для конкретного географическим местоположениям.

Выводы. Результаты проведенного исследования и полученного алгоритма могут быть полезны различным заинтересованным сторонам для принятия решений о актуальных точках в городе, его проблемах и направлении развития.

Список использованных источников:

1. Olieman A., Kamps J., Merino Claros R. LocLinkVis: A Geographic Information Retrieval-Based System for Large-Scale Exploratory Search // 2015. [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1509.02010.pdf.

2. Funkner A.A. Geographical Topic Modelling on Spatial Social Network Data / A.A. Funkner, L.O. Elkhovskaya, I.D. Lenivtceva, M.P. Egorov, A.D. Kshenin, A.A. Khrulkov // Procedia computer science. – 2021. – V. 193. – P.22-31.

3. Zharikov A. Information Retrieval System for News Articles in Russian / A. Zharikov, K. Kristalovsky, V. Pivovarov // Web of Data: The joint RuSSIR/EDBT 2011 Summer School, August 15–19, 2011, Proceedings of the Fifth Russian Young Scientists Conference in Information Retrieval / B. Novikov, P. Braslavsky (Eds.). — St. Petersburg, 2011 — P. 5-14.

4. Jones C.., Purves R.S. Geographical information retrieval // International Journal of Geographical Information Science. – 2008. – V.22. – P. 219-228. [Электронный ресурс]. – Режим доступа: DOI:10.1080/13658810701626343.

Авдюшина А. Е. (автор) Подпись

Бессмертный И.А. (научный руководитель) Подпись