

Использование семантических моделей при мониторинге общественного мнения в Интернет

Ландэ Д.В., Березин Б.А., Павленко О.Ю.

Постановка проблемы. Под семантической моделью (СМ) обычно понимают информационную модель предметной области, имеющую вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги задают отношения между ними. Объектами могут быть понятия, события, свойства, процессы, т.е. семантическая сеть (СС) является одним из способов представления знаний [1]. В составе таких ресурсов, как соц. сети, блоги, каталоги товаров, веб-страницы Интернет-магазинов, новостных агентств и т.д., содержится много сообщений и комментариев, выражающих мнение, отношение, оценку, отзыв и т.д., касающихся явлений, предметов, личностей, тем и т.п. Эта информация, непрерывно создаваемая пользователями Интернета и отражающая общественное мнение по разным вопросам, может собираться системами мониторинга [2], анализироваться и учитываться при планировании деятельности гос. органов, компаний, общественных организаций и т.п.

Использование СМ при мониторинге общественного мнения в Интернет обеспечивает визуализацию тематической структуры текстов, отражающих общественное мнение. Формализация представления таких текстов в виде сетей позволяет применить к ним математические методы и средства анализа. В настоящее время методы автоматизированного построения СМ и их мат. анализа в рассматриваемой предметной области исследованы недостаточно, данная работа направлена на решение этой задачи. В работе предлагается пошаговый метод использования СМ при мониторинге общественного мнения в Интернет.

Анализ публикаций. В работе [3] анализируется отличие между общественным мнением о генетически-модифицированных организмах, представленном в ресурсах Интернет и мнением экспертов в научных публикациях. Для этого рассматривался контент трех групп текстов, содержащих термин “GMO”: статьи с федеральных и регуляторных сайтов США, включая FDA, NSF и т.д.; десять веб-сайтов из списка на первой странице результатов поиска “GMO” в Google; 660 заголовков из Google News найденные по запросу “GMO” за год. Для этих ресурсов были построены три СС на основе анализа смежности слов, в качестве концептов использовались слова с частотой повторения выше среднего значения. В результате были определены наиболее центральные слова в каждой сети, сравнивались слова между сетями, оценивалось, какое построение является более позитивным, негативным или нейтральным.

В работе [4] анализ общественного мнения состоит в выявлении тематической структуры в массиве комментариев при обсуждении резонансного фильма, размещенного на канале Youtube. При этом, сравниваются результаты семантического анализа и тематического моделирования. Для анализа с Youtube было собрано около трех тысяч наиболее популярных комментариев и постов. При построении СМ в качестве вершин использовались биграммы. Тематическая структура обсуждения была выявлена с помощью кластеризации основной компоненты СС. Темы, выделенные при помощи сетевого анализа и в результате тематического моделирования в значительной степени совпали. Сделан вывод, что благодаря визуализации результатов, семантический анализ может дополнять тематическое моделирование либо служить эффективной альтернативой.

В работе [5] анализируется общественное мнение о выборах президента США в 2012 г. на основе новостных статей, опубликованных в Интернет. С помощью системы мониторинга были собраны более 81 тыс. англоязычных статей (январь - июль 2012 г., предварительные выборы) и более 49 тыс. статей (август – октябрь, выборы) с более чем 400 информационных

агентств. На основе этих ресурсов были выделены триплеты субъект-глагол-объект и с помощью их построены, соответственно, два семантических графа, отражающих основные действующие лица, их избирательные лагеря и т.д. Результаты анализа избирательной кампании были получены с помощью вычисления характеристик семантических графов: действующие лица и темы каждого из лагерей, наиболее центральные узлы и центральные роли в каждом из лагерей,

Методы и средства построения семантических моделей. Анализ работ [3-5] позволяет сформулировать следующие особенности построения СМ, используемых при мониторинге общественного мнения. На первом этапе построения СМ выполняется предварительная обработка собранных текстов, отражающих общественное мнение, которая состоит в удалении стоп слов, приведении слов к одной форме (стемминг) и т.д.

На следующем этапе выполняется построение СС, графа. В работе [3] сеть строится на основе матрицы смежности слов. В качестве узлов сети использовались слова с частотой повторения выше среднего значения. В работе [4] также используется матрица смежности слов, но в качестве вершин графа были выбраны биграммы с окном (расстоянием в предложении) равным 4. В работе [5], в отличие от традиционного использования матрицы смежности слов, семантический граф строится на основе выделенных в процессе парсинга триплетов субъект-глагол-объект. После такого выявления понятий и их отношений, строится сеть, в которой узлы представлены понятиями/объектами, а ребра отмечены положительными или отрицательными весами.

На заключительном этапе, для выявления и оценки общественного мнения, построенные СС анализируются с помощью математических методов, определяются наиболее центральные узлы, центральные роли и т.д.

С учетом выше проведенного анализа, предлагается пошаговый метод использования СМ при мониторинге общественного мнения в Интернет. Метод ориентирован на слежение за динамикой изменения объекта, на определение состояния, перспективы и прогноза относительно объекта мониторинга и предполагает циклическое выполнение следующих шагов.

1. Формирование текущего массива связей с объектом мониторинга.
2. Выявление основных концептов.
3. Определение связей концептов.
4. Составление запросов, соответствующих концептам.
5. Постоянная обработка запросов по динамическому массиву.
6. Анализ динамики упоминания концептов и прогноз.
7. Противодействие негативным явлениям и содействие позитивным.
8. Выявление новых концептов из динамических информационных потоков, возврат на шаг 3.

Выводы. Рассмотрены существующие методы использования СМ при анализе общественного мнения в Интернет. Показаны особенности этих методов и средств. Предложен пошаговый метод использования СМ при мониторинге общественного мнения в Интернет.

1. Ланде Д.В., Фурашев В.М. Термінологічна мережева модель як відображення процесів денцентралізації влади в Україні // Інформація і право, 2017. - N 2 (21). - С. 66-71.
2. Ландэ Д.В., Березин Б.А., Додонов В.А. Обзор особенностей и возможности контент-мониторинга национального сегмента сети Интернет // Реєстрація, зберігання і обробка даних, 2016. - Т. 18. - N 3. - С. 20-38.
3. Jiang, K., Anderton, B. N., Ronald, P. C., Barnett, G. A. (2018). Semantic Network Analysis Reveals Opposing Online Representations of the Search Term "GMO". *Global Challenges*, 2(1).
4. Юдина, Д. И., & Дудина, В. И. (2016). Семантическая сеть на биграммах как метод валидации результатов тематического моделирования в социологическом исследовании. *Журнал социологии и социальной антропологии*, 19(4), 71-83.
5. Sudhahar, S., Veltri, G. A., & Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data & Society*, 2(1), 2053951715572916