ПОСТРОЕНИЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫНА ОСНОВЕ АЛГОРИТМОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА И ГРАФОВЫХ БАЗ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Козуб В.Н., Пилецкий И.И.

Пилецкий И.И. – к.ф-м.н., доцент

В данном докладе рассматривается система рекомендаций на основе схожести контента и с учётом реакций других пользователей (лайки, репосты и т.д.). Такая система является более эффективной, чем традиционный подход (фильтрация), благодаря использованию дополнительных метрик при формировании рекомендации. Применение такой системы позволит пользователям находить релевантные материалы, хранящиеся в социальных сетях. Предлагается реализация рекомендательной системы с применением графовых баз данных. Рассматриваются векторы схожести материалов.

1. Разработка модели данных

В последние годы рост популярности социальных сетей породил огромное количество материалов, сгенерированных пользователями и хранящихся в социальных сетях. Важной задачей является реализация рекомендательной системы, которая позволит пользователям быстрее находить релевантные материалы. Одним из вариантов реализации такой системы может быть рекомендация на основе схожести контента с учётом реакций других пользователей (лайки, репосты и т.д.). В то же время социальные взаимодействия могут быть удобно описаны в виде графовой модели данных.

Под рекомендательной системой понимается система для поиска и предсказания материалов, которые могут быть интересны пользователю. Предсказание даётся с определённой точностью и основывается на ряде факторов, рассматриваемых далее в разделе 2.

Под схожестью контента подразумевается некоторая оценка подобия двух материалов, основанная на ряде критериев.

В среде социальных сетей под материалом может подразумеваться сообщение, твит, пост в блоге и т.д. Любой материал может быть охарактеризован в основном тремя элементами:

Внутреннее содержимое материала и внутренние тэги;

Тэги, назначенные пользователем;

Пользовательские взаимодействия с документом.

Под пользовательским взаимодействием подразумевается любое действие, которое пользователь может совершить с материалом, например, просмотр, комментирование, лайк и т.д.

При традиционном подходе индексируется только внутреннее содержимое документа, и этот индекс затем используется для помощи в нахождении документов, релевантных поисковому запросу пользователя. Этот подход до сих пор пользуется популярностью во многих поисковых системах [1].

В данной работе предлагается использовать комбинированный подход при подсчёте схожести материалов, который включает в себя содержимое материала, его тэги, а также все пользовательские взаимодействия с материалом. Эти три фактора рассматриваются как три измерения документа в социальном пространстве (назовём их «Контент», «Тэг», «Взаимодействие»). Каждое измерение несёт в себе различный взгляд на материал.

В «Контенте» смысл материала задаётся его автором. А «Тэг» отражает то, как материал воспринимают пользователи соцсети. Каждый пользователь может предоставить свой, отличный от других взгляд на материал простым действием: установкой тэга. Во «Взаимодействии» смысл материала задаётся активность пользователей соцсети, их действиями по отношению к данному материалу.

При таком подходе могут быть использованы семантические алгоритмы для извлечения иерархий из концептов. Это позволяет отыскивать связи между тэгами, и таким образом обнаруживать скрытые отношения между на первый взгляд несвязанными материалами.

Схема данных для такой модели может выглядеть следующим образом (см. рис. 1):

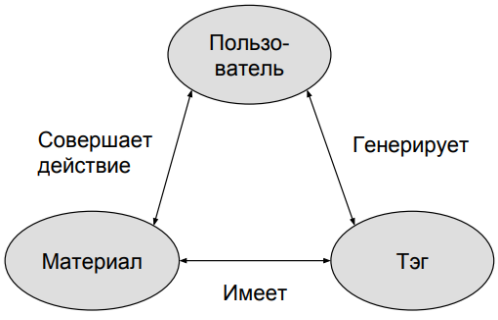

Рис. 1 – Схема данных

На схеме видно, как такая сложная система может быть легко представлена и в дальнейшем расширена при использовании графовых баз данных. Например, база данных Neo4j содержит инструменты для работы с графовым представлением данных [2].

2. Определение схожести контента

Используя собранную о материале информацию, могут быть созданы три различных вектора: вектор контента, вектор социальных тэгов, вектор пользователя.

Вектор контента можно представить следующей формулой:

$$C_i = \{wc(i,1),\ wc(i,2),\ \dots\ wc(i,n)\},$$

где *n* – общее количество тэгов в базе данных, *wc(i,k)* – вес *k*-того тэга в материале или в иерархии тэгов.

wc(i,k) рассчитывается по следующей формуле:

$$a*T(f)-i*df(i,k),$$

где *a* – вес в иерархии, он равен единице, если тэгом является сам материал, иначе – нулю.

Вектор контента представляет собой оценку схожести материалов и учитывает «статичный контент» (информацию в самом материале и внутренние тэги).

Вектор социальных тэгов выглядит следующим образом:

$$T_i = \{wt(i,1),\ wt(i,2),\ \dots\ wt(i,p)\},$$

где *p* – общее количество тэгов в базе данных, *wt(i,k)* – вес *k*-того тэга материала. Таким образом, *wt(i,k)* также является частотой *k*-того тэга в *i*-том документе.

Вектор социальных тэгов представляет собой оценку схожести материалов, основанную на сравнении социальных тэгов материалов, то есть, специальных меток, которые были добавлены потребителями контента, а не его автором.

Вектор пользователя:

$$U_i = \{wu(i,1), wu(i,2),\ \dots wu(i,q)\},$$

где *q* – общее количество пользователей в базе данных, *wu(i,k)* – вес *k*-того пользователя материала. Вес может быть рассчитан различными способами в зависимости от уровня интереса различных пользователей к материалу.

Вектор пользователя представляет собой оценку схожести материалов, основанную на интересе пользователя (его действиях по отношению к материалу).

Также возможно использование более чем одного пользовательского вектора, если необходимо использовать различные веса для различных компонентов (например, один вектор для «лайков», второй для «репостов» и т.д.).

Используя все эти векторы, можно рассчитать различные компоненты схожести, а затем сложить их для получения итогового значения схожести:

$$CombinedSimilarity(i,j) = aCosSim(Ci,\ Cj) + bCosSim(Ti,\ Tj) + c*CosSim(Ui,\ Uj),$$

где *a+b+c=1*.

Стоит отметить, что вычисленная схожесть представляет собой новую информацию, извлечённую из данных в графовой базе данных. Она хранится как модель в рекомендательной системе и может быть использована для предоставления рекомендаций пользователю.

3. Заключение

В данной работе были рассмотрены принципы построения рекомендательной системы для социального контента. Разработаны коэффициенты схожести, основанные на контенте и социальных взаимодействиях. Предлагается их использовать для определения релевантного контента, а также комбинировать их вместе с традиционным подходом (фильтрацией), чтобы получить более релевантные для пользователя результаты. В качестве системы хранения предлагается использовать графовые базы данных, так как модель рассмотренной рекомендательной системы может быть адекватно представлена в виде графа и запросы к данным могут быть выполнены в режиме, близком к реальному времени.

Список использованных источников:

1. Tran Vu Pham, Le Nguyen Thach, “Social-Aware Document Similarity Computation for Recommender Systems”, vol. 00, pp. 872-878, 2011.

2. Neo4jDocumentation [Электронный ресурс] / Neo4j.com – 2018. – Режим доступа: https://neo4j.com/docs/. – Дата доступа: 10.03.2018.