Писъменний І.О. *— рецензент Фіногенов О.Д.*
Інститут прикладного системного аналізу НТУУ "КПІ", Київ, Україна

Кластеризація тегів соціального веб-сервісу

Розглядається задача кластеризації тегів користувачів експертної системи задля встановлення зв'язків між ними. Завдяки об'єднанню тегів у кластери експертна система зможе у відповідь на певний запит пропонувати користувачеві максимально релевантну інформацію, семантично пов'язану з запитом, яка зможе його зацікавити, а отже матиме більший шанс бути проданою.

На даний момент найзручнішим способом відокремлення певної інформації є використання тегів. Цей механізм дозволяє замінити назву об'єкту на його характеристики, завдяки чому для запиту користувачеві достатньо знати лише певні властивості або сферу застосування об'єкту пошуку.

Проте, такий підхід має і недоліки. Основний з них – різні люди описуватимуть один і той самий предмет різним набором тегів залежно від своїх соціальних параметрів, стилю мислення та безлічі інших факторів. Тому постає питання про об'єднання семантично-схожих комбінацій тегів в певні кластери та обчислення вагових коефіцієнтів зв'язків між ними, котрі вказуватимуть їх ступінь спорідненості. Також необхідно враховувати, що один і той самий тег може попасти в різні кластери (для прикладу тег "ягуар" повинен попасти в кластери "авто", "звірі" та "Африка"; тег "дизайн" може стосуватися як графічного, так і веб-дизайну)[2]

Дану задачу можна вирішити алгоритмами, побудованими на теорії графів. Нехай вершинами графу будуть теги. Якщо пара тегів зустрічається в одному запиті, то між ними проводиться ребро, вага якого буде рівною кількості парних зустрічей тегів в запитах. Серед існуючих алгоритмів поставленій задачі найбільш відповідають наступні:[1;4]

1. Кластеризація виділенням зв'язаних тегів: на вході задається мінімальна вага ребра, всі ребра меншої ваги видаляються, після чого залишаються найбільш зв'язані компоненти. Завдяки тому, що в одній вершині може перетинатися декілька ребер ми забезпечуємо можливість використання одного слова в різних контекстах. Сенс алгоритму заключається в пошуку оптимальної мінімальної ваги, яка забезпечить отимальне розбиття вихідного графу[1;2]

2. Пошарова кластеризація – відбувається виділення зв'язних компонент графа на певному рівні відстаней між вершинами. Формується послідовність підграфів вихідного графу, котра відображає ієрархічні зв'язки між кластерами, таким чином алгоритм пошарової кластеризації може забезпечувати як плаский, так і ієрархічний поділ на кластери[3]

3. Алгоритм мінімального дерева покриття – спочатку на графі відбувається побудова дерева мінімального покриття, далі відбувається процедура видалення ребер і розбиття на дрібніші[3]

Висновок. Так як розроблюваний сервіс передбачає постійне навантаження на оновлення та невизначеність кількості кластерів і створення нових, в поточній версії використовується алгоритм кластеризації зв'язаних тегів, котрий дозволяє оновлювати вагу окремо взятого зв'язку та виділення нових.

У подальшому планується оптимізація алгоритму оновлення графу тегів та структури бази даних для забезпечення максимально швидкого оновлення вагових характеристик зв'язків між тегами.

Література. **1.** S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In SIAM International Conference on Data Mining, 2005, **2.** Grigory Begelman, Automated Tag Clustering: Improving search and exploration in the tag space , **3.** Обзор алгоритмов кластеризации данных , http://habrahabr.ru/post/101338/, **4.** Fuzzy C-Means Clustering, http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html.