SWorld – 16-26 December 2014
http://www.sworld.education/index.php/ru/conference/the-content-of-conferences/archives-of-individual-conferences/dec-2014
PERSPECTIVE INNOVATIONS IN SCIENCE, EDUCATION, PRODUCTION AND TRANSPORT '2014
Доклад/ Физика и математика - Информатика и кибернетика

Борозенний С.О.

ПРО ОСОБЛИВОСТІ ВИКОРИСТАННЯ АЛГОРИТМУ LSA

Національний університет «Києво-Могилянська академія»
Київ, Сковороди 2, 04655

Borozennyi S.O.

ABOUT USING THE LSA ALGORITHM

National University of «Kyiv-Mohyla Academy»,
Kyiv, Skovorody 2, 04655

Анотація. В роботі розглядається алгоритм латентного семантичного аналізу та його можливості для моделювання асоціативно-семантичних зв'язків між словами..

Ключові слова: пошук, ЛСА, латентно семантичний аналіз, корреляції між текстами

Abstract. In this paper we describe the the latent semantic analysis algorithm and its ability to simulate associative and semantic relationships between words.

Key words: Search, LSA, latent semantic analysis, correlation between the texts

Вступ

Інтенсивний розвиток інтернету, збільшення кількості інформації яка зберігається в ньому, та яку знаходять саме в інтернеті змушує робити пошук більш інтелектуальним. І найголовніше в цьому пошуку на данний момент, це не збільшення можливих обмежень, а саме інтелектуальність цього пошуку. Тобто пошук не за конкретними словами, а за змістом цих слів та фраз. Створення такого пошуку дасть можливість не тільки шукати інформацію, але й порівнювати зміст текстів, що зробить великий крок к автоматизації популярного на сьогоднішній день електронного навчання.

В більшості випадків пошук відбувається за співпадінням слів, а не за співпадінням змісту цих слів. Тому мінімальні зміни в формулюванні пошукового запиту призводять до суттєвих змін в результаті пошуку, при тому, що сенс пошукової фрази не був змінений.

Латентний семантичний аналіз

Одним з перспективних методів, що дозволяє отримувати дані про значення наведеного тексту, є метод латентного семантичного аналізу (ЛСА).

ЛСА дозволяє виявити значення слів з урахуванням контексту їх використання шляхом обробки великого набору текстів. Принцип дії методу полягає в тому, що порівняння всіх контекстів, у яких слова або групи слів вживаються, і контекстів, у яких вони не вживаються, дозволяє зробити висновок про ступені близькості змісту цих слів чи груп слів.

Вперше метод ЛСА був описаний в роботі "An Introduction to Latent Semantic Analysis" Landauer, T. K., Foltz, P., and Laham, D. у 1998 році [1] і потім розвинений в працях Scott Deerwester, Susan Dumais, George Furnas.

ЛСА використовується для виявлення латентних (прихованих) асоціативно-семантичних зв'язків між термами (словами) шляхом скорочення факторного простору терми-на-документи. Термами можуть виступати як слова, так і їх комбінації, документами - в ідеалі: набори тематично однорідних текстів, або просто будь-який об'ємний текст, довільно розбитий на шматки, наприклад абзаци.

Основна ідея латентно-семантичного аналізу полягає в наступному: якщо у вихідному ймовірносному просторі, що складається з векторів слів (вектор - речення, абзац, документ тощо), між двома будь-якими словами з двох різних векторів може не спостерігатися ніякої залежності, то після деякого алгебраїчного перетворення даного векторного простору ця залежність може з'явитися, причому величина цієї залежності буде визначати силу асоціативно-семантичного зв'язку між цими двома словами.

Наприклад, розглянемо два простих повідомлення з різних джерел:

Перше джерело реклама: «Цей чудовий телефон XXX має потужний акумулятор!»

Друге джерело блоги: «До речі, у девайса XXX непогана батарейка».

Оскільки лексика блогів і реклами не сильно перетинається, то слова «акумулятор» і «батарейка» отримають різну вагу, скажімо, перше маленьку, а друге, навпаки, велику. Тоді ці повідомлення можна об'єднати тільки на основі назви «XXX» (сильний критерій), але подробиці про батарею (назвемо його слабким критерієм) пропадуть.

Однак, якщо ми проведемо ЛСА, то ваги у «акумулятора» і «батарейки» вирівняються, і ці повідомлення можна буде об'єднати на основі хоча і слабкого критерію, але найбільш важливого для товару критерію.

Таким чином, ЛСА виявляє кореляцію між словами різними за написанням, але близькими за змістом.

Висновки

В ході даної роботи було проаналізовано: метод латентного семантичного аналізу (ЛСА); можливість його використання для пошуку документів за пошуковими запитами та встановлення змістових кореляцій між текстами.

Література:

1. Landauer, T. K., Foltz, P., and Laham, D. 1998. An Introduction to Latent Semantic Analysis. DiscourseProcesses, 25: 259-284.
2. Веб-сайт Pearson Knowledge Technologies – http://www.k-a-t.com
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41: 391-407.
4. Foltz, P. W. 1996. Latent Semantic Analysis for text-based research. Behavior

Статья отправлена: 11.12.2014г.