MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
UKRAINIAN FEDERATION OF INFORMATICS

PROCEEDINGS

OF THE IV INTERNATIONAL SCIENTIFIC CONFERENCE

ADVANCED INFORMATION SYSTEMS AND TECHNOLOGIES

AIST-2016

May 25 –27, 2016
Sumy, Ukraine

Comparison and Search of Texts Using Vector Space Model

T. V. Plavin
National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine, taras.plavin@gmail.com

Abstract. The article deals with the issues of coping of information. It outlines one of the technics of a text comparison and search of similar texts. The core logic of this technic is in using of the vector space model. Presented a way of obtaining of a quantitative evaluation of similarity of two texts and finding of matching offers.

Keywords. Text Comparison, Text Search, Vector Space Model, Natural Language, Plagiat.

ВВЕДЕНИЕ

В связи с постоянным увеличением скорости и объёмов публикаций, которое еще носит название «информационный взрыв», все более актуальной проблемой становится проблема оригинальности информации. Она привлекает пристальное внимание как бизнеса, так и научного мира.

Целью доклада является обзор применения модели векторного пространства для сравнения текста и получения количественной оценки сравнения.

МОДЕЛЬ ВЕКТОРНОГО ПРОСТРАНСТВА

Векторная модель – коллекция документов, представленная векторами в одном общем для всей коллекции векторном пространстве. Документ в этой модели представляется как множество термов в неупорядоченном виде. Словом «терм» в информационном поиске обозначают слово, которое составляет часть текста. Документом может быть текст, предложение или другая текстовая единица, используемая для сравнения [1].

Каждый терм имеет свой определенный вес («влияние»), зависящий от количества появлений данного терма в конкретном документе. Существует несколько способов определения веса терма. Среди стандартных функций взвешивания можно выделить такие способы:

1) Булевский вес – ровняется единице, когда терм есть в документе и нулю в ином случае. Данный способ лучше всего использовать для сравнения предложений, так как одинаковые слова редко встречаются в одном предложении, но зато позволяет сократить ресурсы вычисления требуемые для создания векторного пространства [2].
2) Tf (term frequency) - вес задается зависимостью от количества появления терма в документе. Недостатком данного способа является то, что каждый терм считается одинаково важным, а это значит, что служебные слова и слова которые соответствуют тематике текста будут мешать поиску и сравнению [2].
3) Tf-idf (term frequency – inverse document frequency) - вес определяется как произведение функции от количества вхождений терма в документ и функции от величины, обратной количеству документов коллекции, в которых встречается этот терм. То есть он будет максимальным, если терм встречается много раз в небольшом количестве документов и минимальным, если терм встречается почти во всех документах [2].

Для получения векторного пространства все термы, которые содержатся в документах

обрабатываемой коллекции, нужно упорядочить. Для начала нужно создать пространство, размерность которого равна количеству различных терминов во всей коллекции. Затем для каждого документа создать вектор в этом пространстве, учитывая веса термов, и термы которых нету в документе. Размерность как пространства, так и векторов является одинаковой [1], [2].

ПРИМЕНЕНИЕ ВЕКТОРНОГО ПРОСТРАНСТВА В СРАВНЕНИИ ТЕКСТОВ

Рассмотрим простейший пример сравнения двух текстов. Первый текст содержит три слова: «Привет, моряк Иван», а второй текст два слова: «Иван - моряк». Тогда векторное пространство будет состоять из двух векторов в трехмерном пространстве, где каждая ось – это одно уникальное слово (рис. 1). При таком представлении происходит потеря относительного порядка следования терминов, что позволяет находить плагиат, даже при смене порядка терминов.

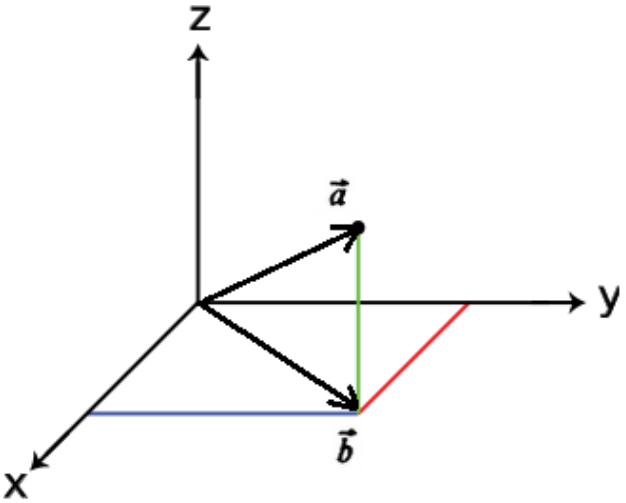

Рисунок 1 – Представление текстов в векторном пространстве

В данном случае слову «Иван» соответствует ось X, слову «моряк» ось Y, а слову «привет» соответственно ось Z. Поскольку слова не повторяются и сравниваются всего два текста, то можно использовать любой метод взвешивания. Для простоты используется булевский метод, поэтому вектор a (текст №1) имеет координаты (1; 1; 1), а вектор b координаты (1; 1; 0). Для определения совпадения текстов используется косинусная мера сходства между векторами (1), которая позволяет компенсировать влияние длины документа (если длина документов сильно отличается). В данном случае мера сходства будет равна 0.82.

$$\cos\alpha = \frac{(\bar{a},\bar{b})}{|\bar{a}||\bar{b}|} \quad (1)$$

В общем же случае векторное пространство может быть представлено формулой (2), где d_j - векторное представление j-го документа, w_{ij} - вес i-го терма в j-м документе, n - общее количество различных термов во всех документах коллекции [1].

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{nj}) \quad (2)$$

В реальной ситуации перед сравнением, тексты нужно обработать: заменить измененные похожие латинские и кирилические символы, использовать для сравнения не слова, а основы слов, убрать служебные символы и тд. тп.

ВЫВОДЫ

Достоинством векторного представления текстов для сравнения и поиска является его простота и достаточно хорошая точность.

Этот метод можно использовать и для сравнения предложений, если мера сходства текстов превышает определенную норму. С чего следует, что он может быть использован в качестве ядра системы поиска плагиата.

REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze An Introduction to Information Retrieval Draft. Online edition. Cambridge University Press. - 2009. - 544 pp.

[2] Daniel Jurafsky, James H. Martin Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition . Pearson Education International . -2009 – 1024 pp.