УДК 004.6

# О ПОДХОДАХ ДЛЯ ОПРЕДЕЛЕНИЯ МЕРЫ НЕСХОДСТВА В ТЕКСТОВЫХ ДАННЫХ

***© 2019 А. Д. Решетников***

*Воронежский государственный университет (Воронеж, Россия)*

*Из-за стремительного роста текстовых данных исключительно важно обрабатывать ее. В статье рассмотрены различные способы получения меры сходства/несходства для текстовых данных. Представлены различные методы учитывающие лексическое сходство строк, а также семантическое расхождение текста.*

*Ключевые слова: мера сходства/несходства, семантическое сходство, лексическое сходство, текстовые данные*

**Введение**

Одна из актуальных сегодня проблем – неостановимое увеличение информации [1]. Одной из ее форм являются текстовые данные. Ручная обработка всего потока данных давно невозможна, однако данный источник используется повсеместно и является очень важным. Измерение сходства текстовых данных может помочь в решении этой огромной проблемы. Расчет сходства между словами – это основной этап для сходства предложений, параграфов и документов. Подход вычисления схожести текста может облегчить людям поиск соответствующей информации. К примеру, это основа для так называемого интеллектуального анализа данных (DataMining), для таких операций как поиск и получение информации (Information Retrieval), классификация текста, извлечение информации (Information Retrieval), кластеризация документов [2], анализ настроений, машинный перевод, обобщение текста и обработка естественного языка (Natural Language Processing).

Лексическое и семантическое сходство слов является существенным элементом сходства предложение, параграфа и документа. Лексическое сходство – это степень того, что две заданные строки схожи посимвольно. Если оценка равна единице (1), это означает, что слова на 100% лексически идентичны. Напротив, ноль (0) означает, что между заданными строками нет общих подпоследовательностей строк. В свою очередь, семантическое сходство представляет сходство между текстами на основе их контекстного значения. Например, пара «дедушка» и «девушка» имеют высокое лексическое сходство, но они не связаны семантически. Пара «автомобиль» и «колесо», которые лексически не схожи, но они связаны семантически, потому что они являются автомобильными терминами.

Таким образом, существуют различные алгоритмы, позволяющие рассчитывать сходство тестовой информации.

**1. Лексическое сходство с помощью алгоритмов на основе строк**

Сходство на основе строк – это наиболее старый, самый простой, но самый популярный метод измерения. Меры на основе строк работают с последовательностями строк и композицией символов. Метрика строки – это метрика, которая измеряет сходство или различие (расстояние) между двумя текстовыми строками для приблизительного сопоставления или сравнения. Двумя основными типами функций сходства строк являются символьные функции сходства (Character-based) и терм функции сходства (Term-based).

**1.1 Меры сходства на основе символов**

Одной из наиболее известных является мера *наиболее длинной общей подстроки*, которая определяет сходство двух строк на основе длиннейшей подпоследовательностей символов, входящих в обе строки.

*Расстояние Дамерау – Левенштейна* определяет расстояние между двумя строками как минимальное количество операций необходимое для превращения одной строки в другую, где операции могут быть вставки, удаления, замены и транспозиции (перестановки двух соседних символов) [3].

Решетников Александр Дмитриевич – Воронежский государственный университет, аспирант кафедры Вычислительной математики и Прикладных информационных технологий, reshetnikov.alex93@gmail.com

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) \text{ если } \min(i,j) = 0 \\ min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} \text{иначе} \end{cases} (1)$$

где $1_{a_i \neq b_j}$ индикаторная функция равная нулю в случае $a_i = b_j$ и 1 в противном случае, и $d_{a,b}(i,j)$ это расстояние между первым $i$ символом в $a$ и первым $j$ в $b$.

*Расстояние Джаро* основано на количестве или порядке символов между двумя строками, которые являются общими [4, 5]. *Сходство Джаро – Винклера* это расширение расстояния Джаро. Чем меньше расстояние Джаро – Винклера $d_w$ для двух строк, тем больше сходства имеют эти строки друг с другом. Результат нормируется, так что $d_w = 0$ означает отсутствие сходства, а $d_w = 1$ – точное совпадение. Сходство Джаро – Винклера равно $1 - d_w$.

$$d_j = \begin{cases} 0 \text{ если } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \text{иначе} \end{cases}, (2)$$

где $|s_i|$ – длинна строки, $m$ – число совпадающих символов и $t$ половина числа транспозиций.

*N-граммы* – это подпоследовательности из n элементов из данной последовательности текста. Алгоритмы подобия N-граммы сравнивают n-граммы от каждого символа или слова в двух строках. Расстояние вычисляется путем деления числа похожих n-грамм на максимальное количество n-грамм [6].

### 1.2 Меры сходства на основе термов

Меры сходства на основе термов, также известные как меры на основе токенов, потому что представляет каждую строку как набор токенов. Эти функции подобия измеряют сходство двух строк на основе общих токенов, из соответствующих им наборов токенов [7]. Такой подход не очень хорошо работает с большими строками, на самом деле для, например, документов в вычислительном плане это дорогой подход, к тому же менее точный. Основной характеристикой сходства на основе токенов является использование пересечений двух наборов токенов в качестве схожести. Пересечение вычисляется на основе точно схожих пар токенов без учета других подобных термов. Такой подход полезен для распознавания перестановок путем разбиения строк на подстроки.

*Расстояние Манхэттена* также известно, как расстояние городских кварталов, метрика такси, расстояние по абсолютной величине, метрика L1 и др. Оно используется чтобы вычислить расстояние, которое будет пройдено, чтобы пройти от одной точки до другой, идя как будто по сетке (квадратами) [8]. Манхэттенское расстояние – это сумма разностей соответствующих компонентов.

$$d_1(p,q) = \|p-q\|_1 = \sum_{i=1}^{n} |p_i - q_i|, (3)$$

где $p, q$ – вектора.

*Косинусное сходство* – это мера сходства между двумя векторами предгильбертового пространства, которая используется для измерения косинуса угла между ними.

$$S_{cos}(d_1,d_2) = \frac{d_1 * d_2}{\sqrt{(d_1 * d_1)}\sqrt{(d_2 * d_2)}}, (4)$$

*Мера Дайса* – удвоенное количество общих термов в сравниваемых строках, деленное на общее количество термов в обеих строках [9].

$$S_{dice}(d_1,d_2) = \frac{2 * d_1 * d_2}{d_1 * d_1 + d_2 * d_2}, (5)$$

*Евклидово расстояние* измеряется путем вычисления квадратного корня из суммы квадратов разностей между двумя элементами векторов.

$$d_{euc}(d_1,d_2) = \sqrt{(d_1 - d_2) * (d_1 - d_2)} (6)$$

### 2. Семантическое сходство

Существует большое количество алгоритмов, которые используются для поиска и анализа информации, однако, особое место занимают те, которые могут обнаружить скрытые закономерности или неочевидные зависимости. Используя семантический анализ текста, мы можем сказать, например, что два текста похожи, даже если эта похожесть выражена косвенно. Или, например, «лыжи» и «автомобиль» по отдельности относятся к разным категориям, но будучи использованы вместе, могут быть интерпретированы в таких категориях, как «спорт» и «отдых».

### 2.1 Сходство, основанное на корпусах

Основанное на корпусе сходство – это мера семантического сходства, которая определяет сходство между словами в соответствии с информацией, полученной из больших корпусов. Корпус – это большая коллекция письменных или устных текстов, которые используются для изучения языка. Многие основанные на корпусе сходства

или меры связанности основаны на концептуальных ресурсах, таких как Википедия.

*Гиперпространственный аналог языка* (Hyperspace Analogue to Language) создает семантическое пространство из совпадений слов [10]. Пословная матрица формируется с каждым матричным элементом являющимся силой ассоциации между словом, представленным строкой, и словом, представленным столбцом. Можно исключить столбцы с низкой энтропией из матрицы. Когда текст анализируется, интересующее слово помещается в начало окна из десяти слов, в котором записывается, какие соседние слова учитываются как схожие. Матричные значения накапливаются путем взвешивания совместного вхождения, обратно пропорционального расстоянию от целевого слова. Считается, что более близкие соседние слова отражают больше семантики целевого слова и поэтому имеют больший вес. ГАЯ также записывает информацию об упорядочении слов, обрабатывая совпадение по-разному в зависимости от того, появилось ли соседнее слово до или после целевого слова.

*Латентный семантический анализ* (Latent Semantic Analysis) является одним из наиболее популярных мер сходства на основе корпусов [11]. ЛСА предполагает, что слова, близкие по значению, будут встречаться в похожих фрагментах текста. Матрица, содержащая количество слов в каждом абзаце (строки представляют собой уникальные слова, а столбцы представляют каждый абзац), построена из большого фрагмента текста, а математический метод, который называется разложением по сингулярным числам, используется для уменьшения количества столбцов уменьшения количества столбцов при сохранении структуры сходства между строками. Затем слова сравниваются путем взятия косинуса угла между двумя векторами, образованными любыми двумя рядами.

*Явный семантический анализ* (Explicit Semantic Analysis) – это мера, используемая для вычисления семантической взаимосвязи между двумя произвольными текстами [12]. Техника, основанная на Википедии, представляет термы (или тексты) в виде многомерных векторов.

Каждая запись вектора представляет вес TF-IDF между термином и одной статьей в Википедии. Семантическая связь между двумя терминами (или текстами) выражается косинусной мерой между соответствующими векторами. TF-IDF (TF – term frequency, IDF – inverse document frequency) – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

### 2.2 Сходство, основанное на знаниях

Меры семантического сходства, использующие информацию из семантических сетей для определения степени сходства слов, называются мерами сходства, основанными на знаниях [13]. Сходство на знаниях состоит из семантического сходства и семантического родства. Семантического сходство определяет два взаимозаменяемых понятия, тогда как семантическое родство связывает понятия семантически. Такой подход использует явное представление знаний, таких как взаимосвязь фактов, значений слов и правил, для описания выводов по конкретным областям. Схема представления знаний обычно включает в себя правила выводов, логические предложения и семантику сети, такую как таксономия и онтология.

Некоторые доступные онтологии: WordNet, SENSUS1, Cyc2 и STDS6 [14]. WordNet является наиболее популярным онтологическим ресурсом и широко используется для измерения сходства на основе знаний. WordNet – это большая английская лексическая база данных исследовательского проекта, разработанного Принстонским университетом. WordNet объединяет существительные, глаголы, наречия и прилагательные в одну концепцию семантических отношений, называемых наборами синонимов (synsets), которые представляют одну концепцию. И концептуально-семантические, и лексические отношения связывают системные сети. Слова в WordNet структурированы иерархически с использованием гипонимии и гиперонимии, и слова могут легко рассматриваться как понятия. Таким образом, WordNet можно интерпретировать как таксономию. Подход, основанный на знаниях, который использует онтологию WordNet, можно разделить на четыре категории: на основе путей, на основе информационного контента (ИК), на основе признаков и других типов.

1) Мера на основе пути. Основным понятием (также известным как меры подсчета

ребер) является длина пути и его положение в таксономии, представленное функцией сходства между двумя понятиями. Эта мера использует кратчайший путь между концепциями.

2) Мера на основе информационного контента. Подход, основанный на ИК, включает в себя определенные концепции в расчете подобия. Основная идея таких мер сходства применяется в модели информационного контекста. Расчет зависит от каждого понятия и потомка частот в текстовом корпусе. Фундаментальная гипотеза должна относиться к более абстрактному понятию с более низкой информацией, а не конкретным содержанием. Подход, основанный на ИК, считается очень перспективным и становится одним из основных направлений исследований в этой области.

3) Измерение на основе признаков. Основная идея семейства сходства на основе признаков заключается в использовании теории множеств между концептуальными наборами функций. Мера на основе признаков описывает набор предполагаемых терминов как свойства или функции. Количество общих характеристик выше, чем менее необычные характеристики двух терминов означает, что эти элементы похожи. Одной из классических мер, основанных на признаках, является модель Тверского [15], которая утверждает, что сходство является антисимметричным. Между функциями подкласса и связанного суперкласса преодолевается вклад его обратного направления в плане оценки сходства.

### 2.3. Смешанные меры сходства

В дополнение к трем категориям, описанным ранее, есть еще несколько мер сходства, которые нельзя отнести ни к одной предыдущей категорий. Идея состоит в том, чтобы объединить ранее описанные подходы, включая основанное на строках, корпусах и основанное на знаниях сходство, чтобы достичь лучшего показателя путем использования их преимуществ.

### Заключение

Как уже говорилось, обработка текстовых данных является важной областью исследований, которая приобретает все большую популярность в последние годы. Измерение сходства между текстовыми документами является важной операцией интеллектуального анализа текста. Было рассмотрено три метода для вычисления сходства для текстовых данных, такие как сходство на основе строк, корпусов и знаний. Меры на основе строк обрабатываются для композиции символов и последовательностей строк. Основанное на корпусе сходство – это семантическая мера сходства, которая определяет сходство между словами на основе информации, полученной из большого корпуса. Семантическая мера сходства, известная как основанное на знаниях сходство, основана на степени сходства между словами и понятиями. Некоторые из этих алгоритмов были объединены во многих исследованиях, и они являются гибридными мерами сходства.

### ЛИТЕРАТУРА


1. Yunianta A. Semantic data mapping technology to solve semantic data problem on heterogeneity aspect / A. Yunianta, O. M. Barukab, N. Yusof, N. Dengen, H. Haviluddin, M. S. Othman // International Journal of Advances in Intelligent Informatics. – 2017. – vol. 3, no. 3. – pp. 161–172.

2. Hidayat E. Y. Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering / E. Y. Hidayat , F. Firdausillah, K. Hastuti, I. N. Dewi, A. Azhari // International Journal of Advances in Intelligent Informatics. – 2015. – vol. 1, no. 3. – p. 132.

3. Hall P. A. V. Approximate string matching / Patrick A. V. Hall, Geoff R. Dowling // Computing Surveys. – 1980. – vol. 12 no. 4. – pp. 381–402.

4. Jaro, M. A. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida / M. A. Jaro // Journal of the American Statistical Society. – 1989. – vol. 84, no. 406. – pp. 414-420.

5. Jaro, M. A. Probabilistic linkage of large public health data file / M. A. Jaro // Statistics in Medicine. – 1995. – vol. 14. – pp. 491–498.

6. Kondrak G. N-gram similarity and distance / G. Kondrak // International symposium on string processing and information retrieval. – 2005. – pp. 115–126.

7. Yu M. String similarity search and join: a survey / M. Yu, G. Li, D. Deng, J. Feng // Frontiers of Computer Science. – 2016. – vol.10, no. 3. – pp. 399–417.

8. Eugene F. K. Taxicab Geometry / F. K. Eugene. – Dover Publications, 1987. – p. 96

9. Dice L. R. Measures of the Amount of Ecologic Association Between Species / L. R. Dice // Ecology. – 1945. – vol. 26, no. 3. – pp. 297–302.

10. Lund K. Semantic and associative priming in high-dimensional semantic space / K. Lund // Proc. of the 17th Annual conferences of the Cognitive Science Society. – 1995, pp. 660–665.

11. Landauer T. K. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge / T. K. Landauer, S. T. Dumais // Psychological Review. – 1997. – vol. 104, no. 2. – pp. 211–240

12. Gabrilovich E. Computing semantic relatedness using wikipedia-based explicit semantic analysis / E. Gabrilovich, S. Markovitch // IJcAI. – 2007. – vol. 7. – pp. 1606–1611

13. Mihalcea R. Corpus based and knowledge-based measures of text semantic similarity / R. Mihalcea, C. Corley, C. Strapparava // American Association for Artificial Intelligence. – 2006. – vol. 6. – pp. 775–780,

14. T. Slimani Description and Evaluation of Semantic Similarity Measures Approaches / T. Slimani // International Journal of Computer Applications. – 2013. – vol. 80, no. 10. – pp. 25–33

15. Tversky A. Features of similarity / A. Tversky // Psychological Review. – 1977. – vol. 84, no. 4. – pp. 327–352, 1977

## ABOUT APPROACHES TO DETERMINE THE MEASURE OF DISSIMILARITY IN TEXT DATA

***© 2019 A. D. Reshetnikov***

*Voronezh State University (Voronezh, Russia)*

*Due to the rapid growth of textual data, it is extremely important to process it. The article discusses various ways to obtain a measure of similarity/dissimilarity for text data. Various methods are presented that take into account the lexical similarity of strings, as well as a semantic discrepancy of the text.*

*Key words: similarity / dissimilarity measure, semantic similarity, lexical affinity, text data*