УДК 004.912

ПОИСК ПОХОЖИХ ДОКУМЕНТОВ НА ПОРТАЛЕ ЯДЕРНЫХ ЗНАНИЙ BELNET

А. П. Дунец

Институт ядерных проблем Белорусского государственного университета, Бобруйская 11, Минск, 220006, Беларусь, E-mail: dunets@gmail.com

В статье рассматривается метод автоматического поиска похожих текстов в системе управления контентом портала. Предложен алгоритм поиска, который основан на вычислении статистических метрик текстов с последующим представлением набора слов в виде числовых векторов. Сравнение текстов по степени похожести производится применением критерия косинусной близости.

Ключевые слова: тезаурус; глоссарий; классификатор; классификация текстов.

Введение

Любая система управления контентом должна обеспечивать средства удобной навигации по множеству документов, которые в эту систему загружены. Во многих существующих решения эта задача целиком ложится на пользователя. Он должен принять решение в какой раздел разместить документ и какие ключевые слова ему присвоить. На практике это требует от пользователя определенных усилий и чревато ошибками и лишними затратами времени. НИИ Ядерных проблем Белгосуниверситета развивается портал ядерных знаний BelNET (Belarusian Nuclear Education and Training Portal, https://belnet.by/) [1]. При разработке портала активно используется подход, основанный на семантических технологиях [2]. В работе [2] рассматриваются понятия «глоссарий» и «тезаурус» и показано как эти элементы семантической теории могут быть применены на практике при создании реальной информационной системы. В то же время в указанной статье конкретные алгоритмы не раскрыты. Данная статья устраняет этот пробел. Рассматривается алгоритма семантической обработки текстов: классификация текстов по степени похожести на основе векторного представления текстов.

1. Векторное представление текстов

Как показывает реальная практика обработки текстов напрямую сравнивать текстовые данные невозможно. Мешает разбиение текста на абзацы, наличие знаков препинания и переноса, наличие разных форм у одного и того же слова. Поэтому необходимо очистить текст от незначащей информации, затем выполнить несколько преобразований полученных данных что бы можно было провести сравнение математическими алгоритмами.

В данной работе использовалась последовательность из следующих преобразований:

1. разделение текста на слова с удалением знаков препинания;
2. преобразование всех слов в нижний регистр;
3. лемматизация слов;
4. расчет частот вхождения слов в тексте;
5. векторизация частотного представления.

Шаги 1 и 2 элементарны и не вызывают вопросов. Остановимся немного подробнее на шагах с 3 по 5.

Шаг 3 лемматизация – преобразование словоформ в нормальную (словарную) форму. Использовались словари библиотеки PHPMorphy [3] и для слов, не найденных в словаре алгоритм Snowball [4].

Шаг 4. Для полученной цепочки слов-лемм вычислялась относительная частота вхождения слов в тексте: отношение числа вхождений слова в тексте к общему числу слов в документе.

Шаг 5. Векторизация. Составлялся словарь лемм для всего корпуса (набора) текстов и для каждого текста формировался вектор частот вхождения лемм из словаря в данный текст.

На простейшем примере это выглядит так. Для двух текстов «Мама мыла раму» и «Раму моют» получаем в результате лемматизации цепочки лемм «мама мыть мама» и «рама мыть» и осуществляем векторизацию, результат которой приведен в табл. 1.

Таблица 1

Простой пример преобразования текста

	Мама	мыть	рама
Текст 1	0,33	0,33	0,33
Текст 2	0	0,55	0,55

Такие числовые представления текстов уже можно обрабатывать математическими алгоритмами.

2. Сравнение текстов по похожести

Современные семантические технологии предлагают множество критериев сравнения векторных представлений текстов. Хорошее сравнение 14 различных метрик приводится в работе [5]. Так же в указанной работе рекомендуется использовать косинусную близость: скалярное произведение векторов.

Для оценки эффективности алгоритма, основанного на критерии косинусной близости, был собран корпус текстов, соответствующий тематике ядерных знаний BelNET. За основу были взяты учебные тексты сайта «Ядерная физика в интернете» [6], нормативные документы сайта Департамента по ядерной и радиационной безопасности МЧС Республики Беларусь [7], материалы BelNET и несколько статей из Википедии по атомной тематике. Всего было проанализировано в сравнении между собой 14 текстов различного размера из одной предметной области и разных источников. Этого оказалось достаточно для вывода о применимости алгоритма.

По оценке алгоритма статьи, из разных источников по одинаковой теме похожи в источнике, но не похожи между источниками. В табл. 2 приведены результаты расчета скалярного произведения векторов для статей «Атом» и «Ядро» сайтов BelNET и Википедия.

Таблица 2

Сравнение статей «Атом» и «Ядро» из разных источников

	BelNET «Атом»	BelNET «Ядро»	Википедия «Атом»	Википедия «Ядро»
BelNET «Атом»	1	0,594	0,071	0,081
BelNET «Ядро»	0,594	1	0,053	0,108
Википедия «Атом»	0,071	0,053	1	0,989
Википедия «Ядро»	0,081	0,108	0,989	1

Глубина и стилистика подачи материала делает тексты очень сильно различными при расчете лингвистических характеристик. Статьи в Википедии написаны для широкого круга пользователей сети интернет, в то же время соответствующие статьи BelNET ориентированы на профильных студентов соответствующих специальностей университета – уровень подачи и стиль материала (насыщенность терминами) существенно отличаются. Это приводит к сильному различию в критерии.

Для текстов с близкими лингвистическими особенностями результаты значительно лучше. Например, если сравнивать текст «Технический кодекс установившейся практики» с текстами «Пояснительная записка к техническому кодексу» и «Постановление МЧС о радиоактивных элементах» получим в первом случае скалярное произведение равное 0.781 во втором случае 0.439. Стилистика документов практически одинакова и уровень подачи одинаков – документы можно сравнивать.

3. Выводы

1. Приведенный в статье алгоритм по результатам тестов показал свою применимость для сравнения текстов по степени похожести.

2. Может использоваться как начальная точка для решения более сложной задачи – классификации текстов. Идея состоит в том, что если какие-то тексты уже размещены на портале и отнесены к определенным разделам, то по результатам сравнения с ними похожий текст так же может быть отнесен к соответствующему разделу.

3. Алгоритм сравнения можно улучшить, добавив шаг удаления шумовых (незначащих) слов из текста. Как правило это предлоги, союзы и т.п. подобные слова. Но составление подобного списка с учетом специфики предметной области является отдельной задачей.

Библиографические ссылки

1. Сытова С.Н. Система управления ядерными знаниями в Республике Беларусь. Журнал БГУ. Физика. 2022. № 2. С. 87-98. DOI: 10.33581/2520-2243-2022-2-87-98.

2. Использование семантических технологий для развития портала ядерных знаний BelNET / С.Н. Сытова [и др.] // Информационные системы и технологии: материалы междунар. науч. конгресса по информатике. В 3 ч. Ч. 3, Респ. Беларусь, Минск, 27-28 окт. 2022 г. / Белорус. гос. ун-т ; редкол.: С. В. Абламейко (гл. ред.). Минск: БГУ, 2022. С. 193-198.

3. phpMorphy morphological analyzer library for Russian, English, German and Ukrainian languages [Электронный ресурс]. URL: https://github.com/cijic/phpmorphy (дата обращения: 29.03.2024).

4. Snowball Stemmers [Электронный ресурс]. URL: http://snowballstem.org/ (дата обращения: 29.03.2024).

5. Bullinaria J. A., Levy J. P. Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study, 2007 Behavior Research Methods, vol. 39, pp. 510-526.

6. Ядерная физика в интернете [Электронный ресурс]. URL: http://nuclphys.sinp.msu.ru/index.html (дата обращения: 29.03.2024).

7. Департамент по ядерной и радиационной безопасности МЧС [Электронный ресурс]. URL: https://gosatomnadzor.mchs.gov.by/ (дата обращения: 29.03.2024).