Ю.Б. Крапивин

К ЗАДАЧЕ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ СХОЖИХ ДОКУМЕНТОВ

Брестский государственный технический университет,
г. Брест, Республика Беларусь
ybox@list.ru

Одной из наиболее важных и трудоемких задач анализа текстовых документов и поиска информации в сети Интернет является проблема выявления схожих документов. Количество задач, в которых требуется решение данной проблемы, достаточно велико: исключение повторяющихся документов из списка результатов запроса поисковой машины, улучшение качества индекса путем устранения избыточности информации, фильтрация спама, построение сюжетов близких, по содержанию новостных сообщений, выявление случаев заимствования без ссылок на авторов.

Существующие подходы к решению данной проблемы могут быть объединены в два класса: использующие технику построения отпечатков документов (fingerprints) и технику ранжирования, разработанную для информационного поиска.

Техника ранжирования состоит из двух этапов: на первом этапе коллекция документов индексируется, на втором этапе к ней отправляется запрос. При этом запрос используется при проведении вычислений оценки подобия для каждого документа коллекции, в соответствии с функцией – мерой подобия. Документы сортируются по убыванию оценки, а документы с наивысшей оценкой возвращаются пользователю. Эта методика не дает четкого ответа «да» или «нет» на вопрос о том, являются ли документы релевантными требованиям пользователя, но упорядочивает их с учетом вычисленной вероятности соответствия. Информация, необходимая для оценки запроса, зависит от используемой меры подобия, изменение которой определяет эффективность ранжирования. Меры подобия, как правило, используют статистическую информацию, о частоте вхождения i-го терма в документ, коллекцию; размере документов, коллекции и др. Наиболее распространенными являются: скалярное произведение, нормализованное скалярное произведение, косинусная мера, мера идентичности.

Другой метод определения схожих документов, использующий технику построения отпечатков документов, основан на работах Manber, Garcia-Molina и Shivakumar. В рамках этого подхода для

каждого документа коллекции строится его компактное описание – отпечаток, fingerprint, представляющий собой некоторый набор идентифицирующих документ контрольных сумм или хеш-значений. При этом на эффективность метода в значительной степени влияет функция, используемая для генерации хеш-значений из текста документа; степень детализации отпечатка, т.е. длина подстроки-отпечатка, извлекаемой из документа; разрешение отпечатка, т.е. число значений, используемых для построения отпечатка; алгоритм выбора подстрок из документа. Дальнейшее сравнение этих отпечатков позволяет определить вероятность подобия документов.

Рассмотренные выше подходы существенно различаются, но все они имеют одинаковые задачи, которые должны быть решены на этапе предобработки документов, предшествующем индексированию. Среди задачами являются: удаление разметки, лишних пробелов, стоп-слов, необходимость учитывать использование различных форм слов, синонимов и перефразированных предложений. Процесс определения схожих документов без решения сформулированных задач будет малоэффективным, поскольку замена слов на синонимичные сократит степень подобия между документами, частичное совпадение многозначных слов может ошибочно ее увеличить, а в случае, если предложения перефразированы даже незначительно, последующее сравнение не даст желаемого эффекта.

В этой связи решение проблемы выявления схожих документов невозможно без лингвистического анализа и, как следствие, использования лексической, лексико-грамматической и синтаксической информации сравниваемых документов. Методы, основанные на работе с синтаксической информацией, применяющие контекстно-независимые грамматики, семантические классы глаголов, используемые для оценки подобия содержания и жанра, определения схожих элементов, даже в перефразированных документах, демонстрируют приемлемые результаты.

В настоящее время из указанного класса быстрое развитие получили системы определения схожих документов, специализирующиеся на выявлении случаев заимствования без ссылок на автора. Среди них можно выделить WCopyfind, CopyCatch, Anti-Plagiat, PlagiatInform.

Однако данные системы оперируют алгоритмами распознавания явного, но не всегда точного, заимствования фрагментов текста, их соответствие по лексическому составу и позициям лексических единиц либо только по лексическому составу, с учетом простейших морфологических преобразований и отношений синонимии. Ни одна из этих систем поддерживает работу только с одним языком, что является их существенным недостатком.

Система должна предоставлять возможность обработки текстовых документов в многоязычной среде, что предполагает существование средств поддержки нескольких языков, а также функциональности машинного перевода текстов. Последнее позволит обнаруживать в анализируемом документе имеющие место заимствования из текстовых документов, представленных как на его языке, так и на других из рассматриваемого множества языков. Кроме того, качественное решение задачи требует разработки и включения в состав системы развитого лингвистического процессора для построения эффективного индекса с целью предварительного выделения из полнотекстовой базы данных части ее документов, наиболее релевантных анализируемому, а также для лингвистического анализа текста с целью распознавания неявного заимствования его фрагментов: с точностью до парадигм лексических единиц и отношений синонимии для них, а также с точностью до перифраз (уровень синтаксического анализа языка).

Таким образом, реализуемая система может быть представлена в виде совокупности подсистем:

лингвистический процессор, обеспечивающий определение предметной области анализируемого документа и выделение из полнотекстовой базы документов того их подмножества, которое соответствует данной предметной области; преформатирование документа; лексический, лексико-грамматический и поверхностно-синтаксический анализ документа; построение индекса документа в виде списка его наиболее информативных терминов; обработку перефразированных предложений, а также словоформ с учетом для них отношений синонимии; построение лингвистического индекса документа;

лингвистическая база знаний, содержащая базовый и вспомогательные словари, классификатор лексико-грамматических свойств, грамматику и распознающие лингвистические модели анализа поддерживаемых языков;

подсистема распознавания эквивалентности текстовых фрагментов двух анализируемых документов, осуществляющая это распознавание на основании лингвистической информации с ее оценкой в виде числового значения вероятности этой эквивалентности;

подсистема машинного перевода текстов, обеспечивающая реализацию функциональности cross-language-поиска;

подсистема пополнения полнотекстовой базы данных другими релевантными текстовыми документами, используемыми на этапе проведения анализа документа.

Подсистема распознавания эквивалентности текстовых фрагментов, оперирующая результатами лингвистической обработки документов, является базовой при решении рассматриваемой задачи.

При её реализации используются алгоритмы, учитывающие [illegible] ческую (о месте и частоте встречаемости), морфологическую [illegible] частью речи являются, в какой форме употребляются и [illegible] информацию о термах анализируемых текстовых фрагментов [illegible] же модификации метода динамического программирования Ньюм [illegible] и Фишера, что, в конечном счёте, обеспечивает высокую эффек [illegible] ность системы в целом.