УДК 681.3

Д.В. Ландэ, dwl@visti.net, А.Т. Дармохвал, hval@visti.net, В.В.Жигало, vladlen@visti.net, hval@visti.net, ИЦ «ЭЛВИСТИ», НТУУ «КПИ»

МАТРИЧНЫЕ КРИТЕРИИ КАЧЕСТВА ВЫЯВЛЕНИЯ ПОДОБНЫХ ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ ПОТОКАХ

Аннотация

Новостные потоки информации, публикуемой на веб-сайтах сети Интернет, включают сообщения, важнейшие и интереснейшие из которых многократно дублируются (в виде перепечаток или содержательных пересказов). Системы автоматического контент-мониторинга, сетевые информационно-поисковые системы содержат отдельные компоненты, предназначенные для определения содержательного дублирования. При этом проблема качества выявления подобных документов при использовании различных критериев является открытой научно-практической проблемой. В данной статье описываются критерии качества выявления подобных документов, основанные на анализе таких свойств так называемой матрицы подобия, как симметричность и транзитивность. Ранее близкие по смыслу критерии рассматривались авторами в работе [1], в настоящей статье представлены более точные и универсальные аналитические выражения для расчета этих критериев, а также приведены результаты экспериментов на многоязычных текстовых корпусах, формируемых с помощью системы контент-мониторинга InfoStream.

Содержательное дублирование новостных сообщений

Задача выявления дублирующихся сообщений (их принято называть дубликатами), а также перепечаток документов с небольшими изменениями («почти дублей») является одной из актуальнейших и сложнейших при интеграции информационных ресурсов. Понятие содержательных дублей документов достаточно расплывчато, авторы даже пытались анализировать такие явления, как пересказ одних и тех же событий, описание различных аспектов разными людьми.

Серьезное упрощение названой задачи может быть получено за счет применения содержательных методов, например, путями ранжирования первоисточников, определения и выделения тематических информационных каналов, экспертного формирования словарей значимых слов и т.п.

Преодоление использования явно дублирующейся информации не представляет проблем, однако дублирующиеся по смыслу сообщения выявляются не так легко, здесь на помощь приходят алгоритмы, аналогичные алгоритмам построения информационных портретов [2], их сопоставления и вероятностной оценки. На практике явные дубликаты выявляются даже с

помощью механизмов контрольных сумм, но этот подход не решает проблем пользователей, для которых чаще всего не имеет значения, с чем они имеют дело, с прямой перепечаткой или с небольшой перефразировкой. Вместе с тем многие недобросовестные издания перепечатывают содержание сообщений, попросту изменяя заглавия (работа хедлайнеров). И такой вид дублирования элементарно обходится с помощью контрольных сумм (но уже без учета заголовков). Дальнейший анализ показал, что при перепечатке материалов чаще всего остаются без изменений несколько первых предложений текста или первый абзац. И этот критерий был учтен и успешно внедрен. Вместе с тем качество выявления содержательного дублирования оставалось недостаточно высоким.

Исследовались методы, основанные на учете повторений встречаемости цепочек слов (например, метод «шинглов» (чешуек), достаточно хорошо описанный в работах [3], [4], [5] и [6]. Этот остроумный и эффективный метод поиска «почти дублей» оказался не очень чувствительным для небольших текстов с возможными перефразировками (авторы с интересом наблюдали эффекты двойного перевода при перепечатках с русского на украинский, а затем снова на русский).

Естественным путем развития исследований стало обращение к статистическим подходам. Еще в 2002 году представители Яндекса опубликовали свою методику выявления дубликатов, основанную на анализе N наиболее «качественных» слов [7]. При этом качество слов определялось экспертами, а соответствующий математический аппарат получил название «нечеткой цифровой сигнатуры». В этом подходе авторов смутил наивный подход, например, при умножениям вероятностей зависимых событий (слов в сообщениях), а также необходимость «ручного» отбора значимых слов (очевидно, важность отдельных слов может изменяться во времени).

В распоряжении авторов был достаточно мощный информационный ресурс одной из служб интеграции новостей - ретроспективная база данных системы контент-мониторинга InfoStream [8]. Система InfoStream применяется для решения задач автоматизированного сбора новостной информации с открытых веб-сайтов и обеспечения доступа к ней в поисковых режимах. Эта разработанная в компании ElVisti система в настоящее время охватывает свыше 2000 источников, а ретроспективные базы данных системы представляют собой корпус объемом более 30 млн. документов. Следует отметить, что процент дублирующихся сообщений в системе InfoStream значительно меньше, чем во всем Интернет-пространстве. Это объясняется подбором источников для сканирования, в число которых входят лишь те, которые хоть изредка публикуют оригинальные материалы.

Алгоритм выявления подобных документов в системе InfoStream

Принцип выявления значимых ключевых слов (далее будем называть их *термами*) базируется на законе Зипфа [8], [9] и сводится к выбору слов со

средней частотой встречаемости (наиболее встречаемые слова игнорируются с помощью «стоп-словаря», а редкие слова из текстов сообщений не учитываются).

Выявление дублирующихся по содержанию новостных сообщений в системе InfoStream выполняется на основе лингвостатистических методов, заключающихся в выявлении в различных документах общих термов, цепочки которых образуют словесные сигнатуры сообщений.

Определение содержательно подобных документов и дубликатов, использующееся в системе InfoStream до настоящего времени, заключается в том чтобы считать дубликатом тот документ, если 6 ключевых слов одного документа совпадают ключевым словам другого документа (из 12 возможных). Следует отметить, что применение более «мягкого» критерия к множеству отобранных термов позволяет реализовать режим «поиска подобных документов». В данном случае ключевые слова (термы) определяются на основании подхода TF IDF [10], полученных при морфологической обработке (стемминга).

Определение дублей документов новым способом заключается в выявлении ключевых слов с использованием морфологических частотных словарей, в которые вошли имена существительные и обще известные фамилии и названия фирм и организаций. Вычисление коэффициентов производится на основании весового подхода - модификации стандартного подхода TF IDF , Okapi BM25 [11]:

$$TF \cdot IDF = \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \cdot \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \ ,$$

где – $f(t, D)$ - частота встречаемости слова q_i в документе D, $|D|$ - длина документа D, $avgdl$ - средняя длина документа в коллекции текстов, общее количество которых - N, $n(q)$ - количество документов в коллекции, содержащих данное слово, k_1, b - параметры, выбираемые экспертами.

Введем обозначения: пусть "$\prec$" – оператор подобия, а "$\equiv$" - оператор дублирования. Очевидно, что для алгоритма выявления подобных документов и дубликатов, о котором идет речь справедливо правило рефлексивности:

$$A \prec A, \quad A \equiv A,$$

где A – произвольный документ.

Оператор подобия не обладает свойством симметричности. Из подобия документа A документу B не следует обратное, т.е.:

$$A \prec B \not\Rightarrow \quad B \prec A.$$

Также не выполняется условие транзитивности:

$$A \not\prec B, \quad B \prec C \not\Rightarrow \quad A \prec C.$$

Действительно, например, отельный документ может быть подобен тексту из подборки, которая его включает, но сама подборка может не быть подобной этому документу. Или документ может быть подобен двум документам, из которых он скомпилирован, но сами оригиналы могут существенно отличаться.

Для отношения дублирования, наоборот, симметричность и транзитивность выполняются:

$$A \equiv B \Rightarrow B \equiv A,$$

$$A \equiv B, \quad B \equiv C \Rightarrow A \equiv C.$$

Заметим, что отношение, обладающее свойствами рефлексивности, симметричности и транзитивности является отношением эквивалентности [9], в нашем случае, отношением содержательного совпадения или дублирования.

Как было замечено свойство дублирования документов является более жестким критерием подобия, например, совпадение 3, 4 или 5 термов свидетельствуют о некоторой содержательной близости, т.е. можно записать:

$$"\prec" \rightarrow "\approx".$$

На практике каждому документу D_i из контрольного документального корпуса по приведенному выше алгоритму совпадения термов в сигнатурах (в разных экспериментах варьировались необходимые количества совпадающих термов) ставился в соответствие вектор с элементами:

$$a_{ij} = \begin{cases} 1, & D_i \equiv D_j, \\ 0, & D_i \not\equiv D_j. \end{cases}$$

Условие симметричности в этих обозначениях записывается следующим образом:

$$\forall \quad i, \quad j \; : \quad a_{ij} = a_{ji},$$

а условие транзитивности:

$$\forall \; i, \; j, \; k \; : \; a_{ij} = 1, \; a_{jk} = 1 \Rightarrow a_{ik} = 1.$$

В соответствии с приведенными рассуждениями были предложены критерии, базирующиеся на вычислении коэффициентов симметричности (S) и транзитивности (T) для матрицы подобия. На контрольном документальном корпусе, изменяя количество сравниваемых в сигнатурах термов, были получены различные значения соответствующих коэффициентов. Коэффициент симметричности рассчитывается следующим образом:

$$S = 2\frac{\sum_{i}^{N}\sum_{j\neq i}^{N} a_{ij}a_{ji}}{\sum_{i}^{N}\sum_{j\neq i}^{N} a_{ij}},$$

а коэффициента транзитивности определяется по формуле:

$$T = \frac{\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{\substack{k\neq i\\ k\neq j}}^{N} a_{ij}a_{jk}a_{ik}}{\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{\substack{k\neq i\\ k\neq j}}^{N} a_{ij}a_{jk}}.$$

где N – количество документов в контрольном корпусе.

Очевидно, что так рассчитываемый коэффициент симметричности ассоциируется с точностью при определении дубликатов документов, а уровень транзитивности – с полнотой.

Вместе с тем следует заметить, что проверка коэффициентов асимметричности и транзитивности может использоваться лишь для формальной проверки приближения отношения к свойствам эквивалентности. Само определение того, что эта эквивалентность – содержательное дублирование было предоставлено аналитиками-экспертами. Приведенный выше алгоритм кроме своего эмпирического подтверждения хорош тем, что позволяет варьировать некоторым числом (количеством сравниваемых термов в сигнатурах), значение которое можно подобрать с учетом оптимизации двух названных коэффициентов.

Экспериментальные данные

Авторами был предложен подход к выявлению дубликатов документов на разных языках, при этом используются переводные эквиваленты ключевых слов, на разных языках, полученных новым подходом к выявлению дубликатов.

Для проведения опыта, был взят массив документов, полученных из системы контент-мониторинга InfoStream, за март 2009 с общем количеством документов более 1,2 млн., доля украинских документов в массиве составляла 146858 и русских 1125359.

Для получения сравнительных данных решено было провести эксперименты по поиску дубликатов с разными данными:

- Поиск дубликатов по ключевым словам, получаемым путем применения стемминга (обрезания изменяемых частей слов), в русско-украинском массиве документов.

- Поиск дубликатов по ключевым словам, получаемым путем использования частотных морфологических словарей .
- Поиск по переведенным ключевым словам.

На рис. 1 и 2 приведены значения коэффициентов симметричности и транзитивности, соответственно.

В таб. 1 показаны результаты исследования массива документов на информационные дубликаты с разным количеством слов для определения дубликатов.

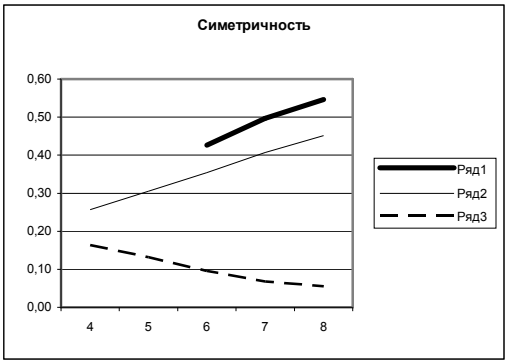

Рис.1. Зависимость коэффициента симметричности от количества совпадающих термов в выявления дубликатов. Ряд 1 – коэф. полученные при выявлении дубликатов (стемминг), Ряд 2 – коэф. полученные при выявлении дубликатов (морфологические словари), Ряд 3 – коэф. полученные при выявлении дубликатов на разных языках

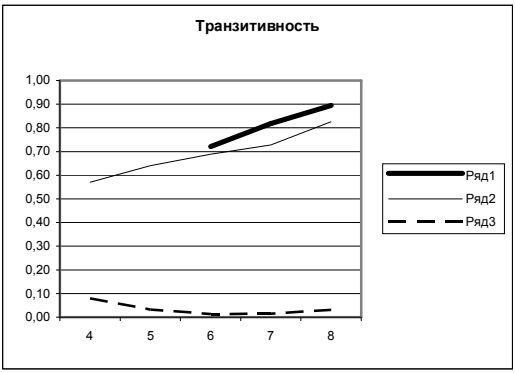

Рис.2. Зависимость коэффициента транзитивности от количества совпадающих термов в выявления дубликатов. Ряд 1 – коэф. полученные при выявлении дубликатов (стемминг), ряд 2 – коэф. полученные при выявлении дубликатов (морфологический словари), ряд 3 – коэф. полученные при выявлении дубликатов на разных языках

Таблица 1. Результаты исследования.

Количество слов	Выявление дубликатов стандартным способом		Выявление дубликатов новым способом		Выявление разноязычных дубликатов	
	Симметричность	Транзитивность	Симметричность	Транзитивность	Симметричность	Транзитивность
4	-	-	0,26	0,57	0,16	0,08
5	-	-	0,31	0,64	0,13	0,03
6	0,43	0,72	0,35	0,69	0,10	0,01
7	0,50	0,82	0,41	0,73	0,07	0,02
8	0,55	0,90	0,45	0,83	0,06	0,03

Таблица 2. Количество найденных дубликатов

Количество слов	Количество дублей стандартный способ	Количество дублей новый способ	Количество разноязычных дубликатов
4	-	23263120	3488705
5	-	9905076	873533
6	11623225	5388866	283027
7	6704983	3412895	125533
8	4525533	2539295	63698

Заключение

Приведенные подходы позволили выбрать наилучшие параметры алгоритма, применяемого в системе контент-мониторинга InfoStream, обеспечивая эффективную селекцию дубликатов, выявление подобных документов в многоязычном текстовом корпусе.

Полученные результаты позволили вплотную подойти к решению проблемы эффективного автоматизированного выявления плагиата в текстах небольших объемов. Эта проблема сегодня имеет большой резонанс [12], [13], но существующие алгоритмы ее решения раскрываются не часто из-за опасений обесценивания наработанных механизмов.

В заключение следует назвать две проблемные области в выявлении дубликатов по представленному алгоритму. Во-первых – это некорректная во многих случаях работа с короткими сообщениями, зачастую вырождающимися в один лишь заголовок. Выявление значимых слов в таких сообщениях - открытая проблема, актуальная, например при реализации метапоисковых систем в Интернет, где приходится иметь дело лишь с короткими рефератами (сниппетами) документов. Вторая проблема связана с длинными документами, обзорами, дайджестами. Термы в словесных сигнатурах таких документов могут не отражать контента каждой составляющей обобщенного документа, а иногда охватывать значительную часть словаря соответствующего языка.

Наряду с вышесказанным, необходимо заметить, что устранение дублирующихся сообщений в информационных потоках требуется далеко не всегда. Существует ряд задач, в которых используется факт дублирования текстов сообщений в различных источниках, например при определении важности сообщения (если сообщения многократно дублируется на сайтах и в СМИ) или при определении эффективности PR-кампаний (подсчет републикаций пресс-релизов и др.)

Литература

1. Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды Восьмой Всероссийской научно конференции RCDL'2006, Суздаль, Россия, 2006. – С. 115-119.
2. Григорьев А.Н., Ландэ Д.В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Труды международной конференции "Диалог'2005". -С. 109-111.
3. S. Ilyinsky, M. Kuzmin, A. Melkov, I. Segalovich. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW2002, 2002.
4. U. Manber. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp. 1-10, January 1994.
5. Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse. Syntactic Clustering of the Web // WWW6, 1997.
6. Andrei Z. Broder. Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – 2000. p 1-10.
7. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. - М.: "Вильямс", 2005. - 272 с.
8. Сегалович И.В. Как работают поисковые системы. // Мир Internet. – 2002. -№ 10.
9. Шрейдер Ю.А. Равенство, сходство, порядок. - М.: "Наука", 1971. - 256 с.
10. Salton G, Buckley C. Term-Weighting Approaches // Automatic Text Retrieval. Information Processing and Management, 1988. - № 24, 5. - pp. 513-523.
11. Lande D.V., Zhygalo V.V. About the creation of a parallel bilingual corpora of web-publications, Publication: eprint arXiv:0807.0311.
12. К. Нейл, Г. Шанмагантан. Web-инструмент для выявления плагиата. // Открытые системы. -2005. -№ 1.
13. Ланде Д.В., Жигало В.В. Підхід до рішення проблеми пошуку різномовного плагіату // Сб.наукових праць "Проблеми автоматизації та управління". - Київ: НАУ, 2008. -Вип. 2(24). -С. 125-129.