Подход к выявлению дублирования сообщений в новостных информационных потоках*

© Д.В. Ландэ, А.Т. Дармохвал, А.Ю. Морозов

Информационный центр «ЭЛВИСТИ»
dwl@visti.net, hval@visti.net, alex@ visti.net

Аннотация

Рассматривается явление содержательного дублирования сообщений в новостных информационных потоках, порождаемых публикациями на веб-сайтах сети Интернет. Представлен критерий выявления дубликатов, а также статистика дублирования информации, сканируемой системой контент-мониторинга InfoStream.

1. Проблема дублирования новостных сообщений

Сегодня Интернет-пространству присущи такие недостатки, как ограниченность интегрированного доступа к информационным ресурсам, обилие «информационного мусора», невозможность гарантирования целостности документов, практическое отсутствие возможности смыслового поиска [1]. Эти проблемы обуславливаются несколькими причинами, среди которых можно назвать непропорциональный рост уровня информационного шума и многократное дублирование информации.

Важные сообщения многократно дублируются на экспоненциально растущем количестве сайтов, в то время, как количество заслуживающих внимания источников растет не такими большими темпами, скорее всего, линейно. Дело в том, что серьезные источники информации - это объекты реальной жизни, в то время как сайты в своей совокупности представляют виртуальное пространство, которое развивается по собственным законам.

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.

* **Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.**

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).

2. Новостная среда

Изначально 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.

Обработка входных сообщений в системе контент-мониторинга InfoStream, вплоть до выявления значимых ключевых слов представлена на рис. 1.

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).

В качестве некоторых «инвариантов» для отдельных сообщений в системе InfoStream используются цепочки из 12 наиболее весомых с точки зрения лингвостатистических критериев термов, прошедших процедуру морфологической обработки (стемминга). Такое небольшое количество термов в цепочке, которая является своеобразной словесной сигнатурой, объясняется небольшой средней длиной новостных сообщений (2000-3000 символов)

Рис. 1. Схема обработки входных сообщений

3. Алгоритм выявления дублей

Итак, 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.

Метод, предложенный авторами и используемый в системе InfoStream, заключается в признании документов дубликатами, если в их сигнатурах совпадает более 5 термов (из 12 возможных). Следует отметить, что применение более «мягкого» критерия к множеству отобранных термов позволяет реализовать режим «поиска подобных документов».

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:

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

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:

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

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

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:

$$"\prec" \quad \rightarrow \quad "\equiv".$$

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:

$$a_{ij} = \begin{cases} 1, & D_i \equiv D_j, \\ 0, & \text{иначе.} \end{cases}$$

Условие симметричности в этих обозначениях записывается следующим образом:

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

а условие транзитивности:

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

Автор исследовали критерии подобия (изменяя количество сравниваемых в сигнатурах термов), чтобы достичь на контрольном документальном корпусе максимального уменьшения коэффициента асимметричности:

$$\frac{\sum_{i}^{N}\sum_{j}^{N} |a_{ij} - a_{ji}|}{\sum_{i}^{N}\sum_{j}^{N} a_{ij}},$$

и увеличения коэффициента транзитивности:

$$\frac{\sum_{i}^{N}\sum_{j}^{N}\sum_{k}^{N} a_{ij} a_{jk} a_{ik}}{\sum_{i}^{N}\sum_{j}^{N} a_{ij}},$$

где N – количество документов в контрольном корпусе.

Очевидно, что так рассчитываемый коэффициент асимметричности ассоциируется с огрублениями при определении дубликатов, 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.

4. Экспериментальные данные

На рис. 2 и 3 приведены экспериментально полученные значения коэффициентов симметричности и транзитивности в зависимости от количества учитываемых совпадающих термов при попытках выявления дублирования. При определении оптимального количества термов, необходимого для выявления дублирования учитывался баланс этих коэффициентов. Кроме того, параллельно результаты оценивались экспертами. Следовало бы отдельно остановиться на субъективном факторе, присутствующем при экспертной оценке уровня дублирования. Сегодня учет этого фактора такая же сложная и неоднозначная задача, как и задача определения пертинентности результатов поиска [8].

Опыт показал, что в русско- и украиноязычных потоках новостей совпадение хотя бы 6 термов в сигнатурах документов приводят к более чем 90% полноты и 95% точности при выявлении содержательных дубликатов.

В соответствии с этим критерием авторами было проведено исследование соотношения дублирующихся и оригинальных сообщений в новостных информационных потоках. Исследования привели к удивительному результату. Оказалось, что количество оригинальных сообщений и их содержательных дублей, охватываемых системой InfoStream в 2005 году, почти в точности совпало (рис. 4).

Это же соотношение справедливо для отдельных событий, отражаемых в электронных СМИ (рис. 5). Лишь некоторые «феноменальные» публикации, дублируются десятки раз.

Авторами также исследовался уровень дублирования для новостных документов, имеющих контекстные ссылки на другие сайты-источники.

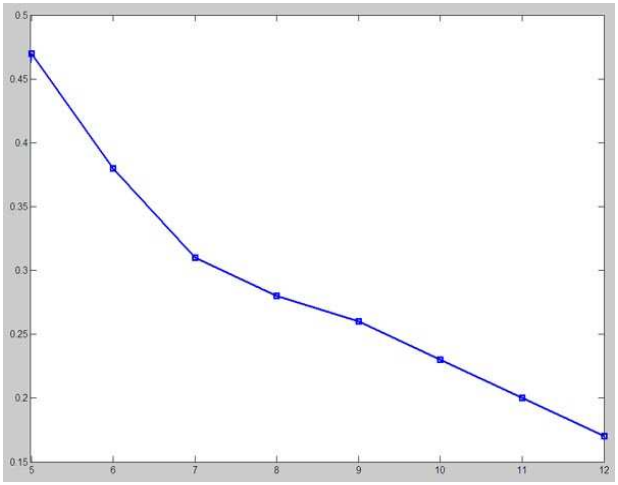

Рис. 2. Зависимость коэффициента асимметричности от количества совпадающих термов в критерии выявления дубликатов

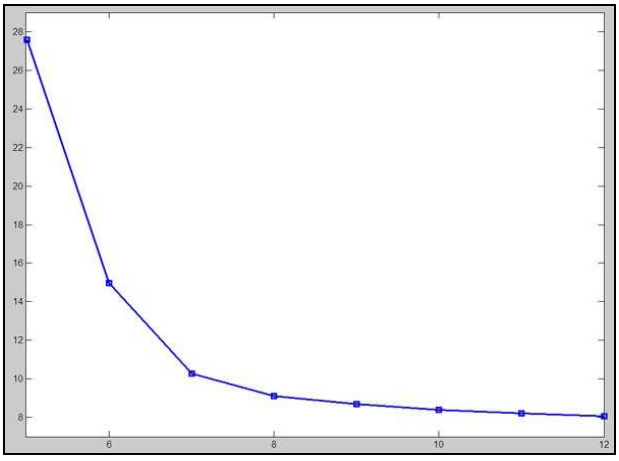

Рис. 3. Зависимость коэффициента транзитивности от количества совпадающих термов в критерии выявления дубликатов

На рис. 6 приведен график зависимости уровня дублирования для источников (исследовалось около 1500 сайтов), ранжированных по количеству исходящих ссылок. По графику видно, что до определенного значения (порядка 800) уровень дублирования значительно превышает средний, равный ~ 50%. При небольшом количестве исходящих ссылок этот уровень понижается, однако при минимальном количестве ссылок снова возрастает. Можно считать, что значения рангов источников 1400 и выше соответствуют «зоне массового плагиата» (ссылок мало, а уровень дублирования - высокий).

Заключение

Проведенные исследования позволили на новом уровне реализовать информационное обслуживание пользователей системы контент-мониторинга InfoStream, обеспечивая селекцию дубликатов. Кроме того, авторами был составлен список наиболее оригинальных информационных источников, сканируемых системой InfoStream, который представляет безусловный интерес для корпоративных пользователей.

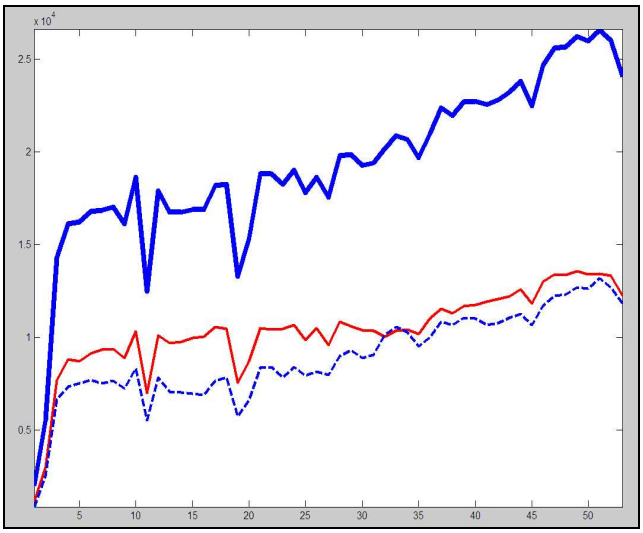

Рис. 4. Объемы информации, сканируемой системой InfoStream в 2005 году, в разрезе недель: сплошная жирная линия – общий объем сообщений, сплошна тонкая – оригинальные сообщения, пунктирная – информационные дубли

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.)

Полученные результаты позволили вплотную подойти к решению проблемы эффективного автоматизированного выявления плагиата в текстах небольших объемов. Эта проблема сегодня имеет большой резонанс [10], [11], 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 - проблема, не решенная авторами до сих пор.

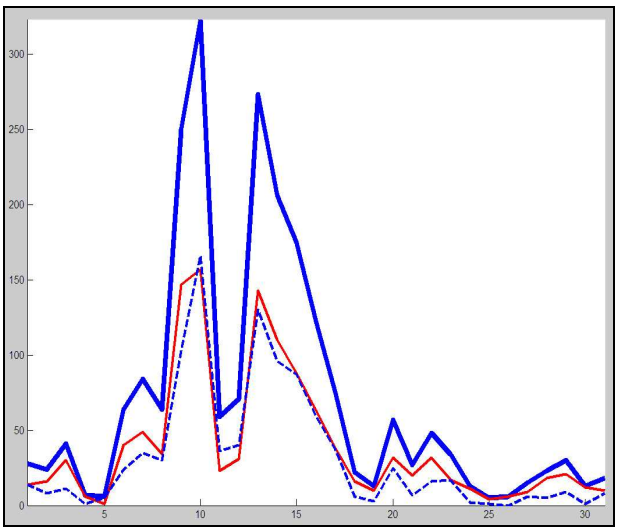

Рис. 5. Объемы информации, сканируемой системой InfoStream в марте 2005 года по запросу «Cebit».

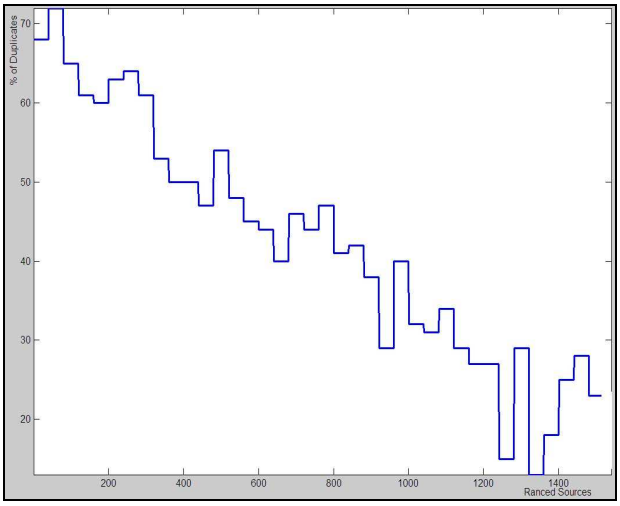

Рис. 6. Уровень дублирования в зависимости от ранга источников

Так два сообщения с результатами футбольного матча на 10 и 40-минуте могут практически не отличаться по набору слов, разница будет лишь в счете. Вторая проблема связана с длинными документами, обзорами, дайджестами. Термы в словесных сигнатурах таких документов могут не отражать контента каждой составляющей обобщенного документа. Обе названные проблемы случая ведут к понижению полноты и точности при выявлении дубликатов и могут рассматриваться как открытая тема для дальнейших исследований.

Литература

[1] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1, - 2005. - № 11. - С. 21-33

[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] К. Нейл, Г. Шанмагантан. Web-инструмент для выявления плагиата. // Открытые системы. - 2005. -№ 1.

[11] W.R. Stone. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! // IEEE Antennas and Propagation, Aug. 2003. -Vol. 45. -№ 4.

The approach to duplication detection in news information flows

D. Lande, A. Darmokhval, A. Morozov

The phenomenon of substantial duplication of documents in the news information flows generated by publications on websites is considered. The criterion of duplication detection in InfoStream system, and also statistics of duplication of the information is submitted.