УДК 004.89:004.4+16.31.21+20.23.19

А. И. Парамонов

«УМНЫЙ ПОЧТАЛЬОН» – АВТОМАТИЗИРОВАННАЯ СИСТЕМА СБОРА НОВОСТНЫХ ЛЕНТ

ВВЕДЕНИЕ

Одна из проблем нахождения информации в сети Интернет обусловлена несовершенными и разрозненными форматами представления информации. На текущий момент данные, распространяемые по сети, наиболее часто представлены в следующих форматах: HTML, PDF, Office (DOC, XLS), Flash (SWF), XML. Каждый формат имеет недостатки в плане анализа информации. Поиск одновременно простого, функционального и унифицированного стандарта привел к созданию RSS-формата. RSS – это основанный на XML формат, предназначенный для сбора информации [1]. В RSS транслируется только полезное содержание без каких-либо элементов дизайна, навигации. Использование XML, формата гибкого и расширяемого, позволяет легко приспосабливать RSS под самые разнообразные задачи. Самой востребованной оказалась задача передачи динамического контента (например, последние новости, курсы валют или сводки погоды), и применение в этих целях RSS-потока оказалось самым подходящим решением, гораздо более мощным и гибким, чем традиционные и давно привычные информеры. Для работы с новостными лентами разработаны два вида программного обеспечения: службы синдикации и службы агрегации. Если службы синдикации нацелены на сбор и доставку информации в единое хранилище, то службы агрегации используются для чтения этих информационных потоков. RSS-агрегаторы - специальные программы, с помощью которых пользователи могут получать доступ к данным из информационных ресурсов в формате RSS. Программное обеспечение по обработке новостных лент в формате RSS (агрегаторы) решают проблему поиска и доставки информации, но оставляют без должного внимания такие задачи, как обобщение данных, их обработку и анализ, то есть не уделяют внимание содержанию загруженной информации. Это влечет большую избыточность обрабатываемых данных. Серьезной проблемой остается многократное дублирование сообщений в информационных потоках. И с каждым

ISSN 1562-9945

днем эта проблема приобретает все большие масштабы. Ведь, как известно, сообщения многократно дублируются в экспоненциально растущем количестве сайтов, в то время, как количество заслуживающих внимания источников растет линейно. Очень часто различные ресурсы содержат один и тот же контент. Учесть явно дублирующуюся информацию не представляет проблем, однако дублирующиеся по смыслу сообщения выявляются не так легко, и здесь на помощь приходят алгоритмы сопоставления контента, сравнения и вероятностных оценок.

На сегодняшний день разработано крайне мало программ предоставляющих расширенный инструментарий по анализу и обработке новостных лент. Существуют множество наработок, подходов и решений по данной проблеме, однако данная область еще должным образом не формализована. Таким образом, разработка интеллектуальных автоматизированных программ для получения информации из новостных лент в настоящее время особенно актуальна.

В качестве решения данной проблемы предлагается метод интеллектуальной фильтрации информационных потоков и разработанная на его основе автоматизированная система сбора новостных лент «Умный почтальон».

Метод интеллектуальной фильтрации новостных сообщений

Идея метода состоит в сравнении текстов сообщений и определении степени их схожести с использованием алгоритмов анализа строк [2,3]. Прямой подход к решению этой задачи сводится к разбиению сообщений на предложения и сравнению множеств, полученных при разбиении. То есть, сравнение методом полного перебора, каждого предложения с каждым. При больших объемах информации данный подход в лучшем случае будет работать долго, в худшем – может вообще не выполниться. Необходимо реализовать подход, при котором возможно сравнение сообщений за допустимое время. Для этого предлагается выделить для сравнения из текстов лишь предложения со схожими параметрами.

Вводится понятие фантом. Фантом – это часть слова, полученная в результате простого отсечения заданного числа символов от конца слова. Число отсекаемых символов не вычисляемое, а задается как внешний статический параметр. Введение понятия «фантом»

позволяет без дополнительных синтаксических и морфологических анализов определить схожесть слов в различных формах. А также исключает из сравнения слова маленькой длины (в основном это, предлоги, местоимения), которые несут второстепенную смысловую нагрузку.

Для описания метода вводятся дополнительные параметры:

A1 – число отсекаемых символов (внешний статический параметр),

A2 – длина фантома,

N1 – число элементов в предложении (число фантомов),

M1 – максимальная длина фантома в предложении (M1 = max{A2}),

M2 – минимальная длина фантома в предложении (M2 = min{A2}).

Процесс работы алгоритма можно описать следующей последовательностью действий:

1. задается значения A1 пользователем (или системой по умолчанию);
2. рассчитаются значения A2;
3. рассчитывается число элементов в предложении (N1);
4. рассчитываются значения M1 и M2 – минимальная и максимальная длины фантомов в предложении;
5. выделяются предложения для обработки. Выборка происходит по рассчитанным для каждого предложения параметрам: N, M1, M2.
6. затем вычисляется функция похожести выделенных предложений.
7. осуществляется фильтрация схожих текстов.

Пусть необходимо сравнить два предложения. Относительно небольшое число элементов в предложении (до 10) и небольшая длина «фантомов» слов позволяют вычислить матрицу их похожести методом «грубой силы», то есть непосредственно вычислить функцию похожести для каждой уникальной пары элементов из двух предложений. Таким образом, функция похожести *i* элемента одного предложения с *j* элеметом второго предложения ($F_{eq}(i, j)$) запишется в виде:

$$F_{eq}(i, j) = \frac{Q_{eq}}{M_{eq}} \times 100\,\%;$$

$$M_{eq} = \max(A2_i, A2_j); \qquad (1)$$

$$i \in [1,m], j \in [1,n],$$

где Q_{eq} - число символов совпадающих в сравниваемых элементах,

M_{eq} - длина большего из двух сравниваемых элементов, m – число элементов в первом предложении, n – число элементов во втором предложении.

Значения функций похожести формируют матрицу схожести предложений A, в которой элемент $a(i,j)$ равен значению функции похожести между элементом i первого предложения и элементом j второго предложения.

Для получения окончательного результата нужно обработать матрицу похожести. Одним из вариантов такой обработки является вычисление стоимости оптимального покрытия. Покрытием матрицы A размерности [m,n] будем называть множество, состоящее из max(m,n) элементов матрицы, взятых из разных строк и разных столбцов. Таким образом, в покрытии матрицы нет двух элементов в одной строке или в одном столбце. Стоимостью покрытия матрицы похожести будем называть сумму значений элементов покрытия. Будем называть покрытие оптимальным, если его стоимость максимальна среди всех возможных покрытий.

В отличие от расстояния редактирования стоимость покрытия «нечувствительна» к перестановкам слов. В некоторых случаях это слишком «огрубляет» функцию сравнения. Для устранения этой ситуации вводится «штраф» для слов, находящихся далеко друг от друга.

Важный вопрос, который возникает при вычислении оптимального покрытия — эффективность. Если в предложении небольшое число слов (до 10 элементов), то для расчета функции похожести можно использовать переборные методы. Для больших матриц необходимо использовать эмпирические методики сокращения перебора.

Для наилучшей оптимизации решения предлагается:

 ISSN 1562-9945

1. Выбирать небольшие значения штрафа за перемещения и перестановку слов.
2. Установить большой «порог» минимально допустимого ненулевого значения функции похожести слов.

Подобный подход позволяет значительно уменьшить число ненулевых значений матрицы, а, следовательно, и число шагов перебора.

Степень похожести двух предложений i и j предлагается вычислять по следующей формуле:

$$\begin{aligned} &F_{eq}^{sent}(i,j) = \frac{MC}{M_{eq}^{sent}} \times 100\% ; \\ &M_{eq}^{sent} = \max(N1_i, N1_j); \\ &i \in [1,k],\ j \in [1,l], \end{aligned} \qquad (2)$$

где MC - максимальное покрытие матрицы значений похожести слов, M_{eq}^{sent} - число элементов в более длинном из сравниваемых предложений, k – число предложений в первом сравниваемом фрагменте текста, l – число элементов во втором тексте.

Конечным результатом сравнения предложений является нахождение метрики μ, которая численно показывает степень схожести двух сообщений.

Данная метрика вычисляется по формуле:

$$\begin{aligned} &\mu = \frac{2 \times S_F}{N_F}; \\ &S_F = \sum_{i,j}^{N_F} F_{eq}^{sent}(i,j); \\ &N_F = \max(k,l), \end{aligned} \qquad (3)$$

где S_F – сумма всех вычисленных значений похожести предложений,

N_F – количество предложений в большем фрагменте текстае.

Таким образом, μ - это среднее значение функции похожести среди всех обработанных предложений. Данная метрика является основным критерием отбора в предлагаемом методе фильтрации.

Система сравнивает полученное значение с некоторым определенным пороговым значением α. Экспериментальным путем пороговое значение α для разрабатываемой автоматизированной системы было установлено равным 70.

В результате предложенный метод позволяет определить схожесть текстов, которые содержат одинаковые блоки информации в разной последовательности.

Автоматизированная система «Умный почтальон»

Автоматизированная система (АС) сбора новостных лент представляет собой RSS-агрегатор с расширенными функциональными возможностями. Основной особенностью системы является извлечение информации из сети Интернет с использованием метода интеллектуальной фильтрации.

АС состоит из двух частей: интерфейсная часть (программа, отвечающая за работу с пользователями) и монитор системы (агент, отвечающий за обновление информации и фильтрацию новостных потоков).

Интерфейс программы позволяет средствами пользовательских настроек устанавливать значения параметра A1, интервала обновления новостей и других параметров обновления (например, выбор интересующих адресов ресурсов). Здесь же в интерфейсной части можно получить доступ к загруженных новостным лентам, уже обработанным и отфильтрованным.

Модуль агента с заданным интервалом периодичности обращается к базе данных и по заданным параметрам загружает новости из сети Интернет. При загрузке каждой новостной ленты она подвергаетсяя обратботке. Агент производит фильтрацию на основе предложенного метода. Монитор системы предоставляет информативные сообщения про обновления в базе. Здесь можно просмотреть какие новостные сообщения сравнивались, значение их схожести, найденное разработанным алгоритмом, а так же количество новостей вновь поступивших и добавленных в базу. По окончанию загрузки программа предоставляет пользователю обновленные данные.

Для работы с данными АС «Умный почтальон» предоставляет пользователю набор стандартных функциональностей программы-агрегатора.

 ISSN 1562-9945

Выводы

Разработанный метод на достаточном уровне решает задачу интеллектуальной фильтрации информации, поступающей из сети Интернет. Он позволяет избавить пользователя от дополнительного и избыточного анализа требуемых данных, тем самым уменьшая затраты ресурсов и времени. Применение метода в системах автоматизации бизнес-процессов позволит улучшить качество обрабатываемой информации и увеличить скорость ее обработки. Система «Умный почтальон» прошла пробные испытания и может быть использована в качестве персонального агрегатора. Результаты полученные при аппробации АС «Умный почтальон» могут быть использованы при построении других систем обработки текстовой информации, а сама система может выступать в качестве подситемы работы с Интернет ресурсами в сложных системах.

ЛИТЕРАТУРА

1. Mark Pilgrim Что такое RSS? (перевод: Александр Качанов), (Документ WWW), URL: http://www.webmascon.com/topics/technologies/9a.asp, 11.05.2008
2. Левенштейн В.И. "Бинарное кодирование удалений, вставок и замен" Доклады академии наук СССР, выпуск 163.
3. Simon Harris and James Ross, Beginning Algorithms / Published 2006 by Wiley Publishing, Inc., Indianapolis, Indiana, 564 pages.