Модернизация расчета центроидов в алгоритме CMU

© Вечур А. В.

Харьковский национальный университет радиоэлектроники
vechur@yahoo.com

© Суяргулова Е. Б.

Харьковский национальный университет радиоэлектроники
s_yv_b@mail.ru

Аннотация

В статье приводится анализ программной системы предназначенной для выделения из общего потока web-новостей кластеров новостей относящихся к одному событию, а так же к одному новостному сюжету. В статье приведены основные идеи, положенные в основу системы, общий алгоритм работы системы, а так же результаты ее тестирования.

1. Введение

На протяжении своего существования Интернет постоянно становился все более доступным и востребованным для постоянно расширяющегося круга пользователей. Одним из переломных этапов в развитии Интернета было появление гипертекстовых технологий названных в последствии Web 1.0, которое было призвано решить появившуюся в 80 – 90-е гг. XX века проблему систематизации содержимого Интернета. Но уже к 2001 году на смену переставшей удовлетворять потребности пользователей Web 1.0 пришла технология Web 2.0, основными принципами которой было повторное использование ресурсов Интернет, в результате чего облегчалось создание новых ресурсов. Одной из особенностей Web 2.0 является появление реализуемых с помощью

RSS-лент web-потоков, которые обеспечивают распространение по сайтам постоянно обновляемой информации [5].

В последние годы на фоне большой мощности и обширности Интернет-ресурсов все более остро встает вопрос о навигации внутри Сети. Частично из-за объемов информации хранящейся в Интернете, частично из-за дублирования одной и той же информации разными сайтами (что связано и с использованием RSS-лент) поисковые системы все чаще выдают слишком обширные для конечного пользователя результаты поиска. Для решения этой проблемы применяют такие подходы как TextMining, и автореферирование которые заключаются в попытке программного анализа смысла текста содержащегося web-ресурсом. Таким образом, развитие Интернета дошло до той стадии, когда актуальной проблемой является облегчение поиска нужной информации. В качестве частного случая такой информации выступает информация позволяющая проанализировать интересующие пользователя события. Одним из источников, которой являются web-новости [5].

Целью данной работы является анализ алгоритма призванного производить анализ web-новостей позволяющий отделить ту часть новостей, которая посвящена одному событию.

2. Поход к исследованию web-новостей основанный на изменении смыслового содержания web-новостей посвященных отдельным событиям

Уже в 90-х годах решалась проблема отделения новостей посвященных одному событию. Для этой цели было разработано целое направление алгоритмов TDT (Topic Detection and Tracked) которые решали эту задачу на основании тематической и временной близости новостей к сформированным подборкам, а также того из каких новостей сформированы эти подборки [1, 2].

Это направление алгоритмов продолжает развиваться и в настоящее время. Естественным развитием алгоритмов TDT в свете глобальной ориентации систем обработки содержимого Интернета на автоматический анализ смыслового содержания, является, использование автоматического анализа смыслового содержания новостей. Примером такого использования является привлечение алгоритмов TextMining для поиска в тексте новости ответов на

вопросы: «где и когда произошло описываемое событие?», «кто был вовлечен в описываемые события?», «каким образом произошло событие?», «какое воздействие, значение, или какие последствия возымело событие на определенные слои населения» [3]. Ответ на последний вопрос чаще всего является целью пользователя читающего новость. А ответы на первые три вопроса бывают, полезны также и при подборе линеек новостей посвященных отдельным событиям (т.к. новости, в которых ответы хотя бы на один из них не совпадают, описывают разные события).

С помощью ответов на приведенные вопросы можно достаточно точно разделить web-новости в соответствии с освещаемыми событиями, но в этом случае проблема состоит в том, что русский язык слабо поддается формализации и поэтому получить абсолютно точные и полные ответы на сформулированные вопросы практически не возможно. В этом контексте менее точными, но более реалистичными являются алгоритмы, основанные на анализе текста как набора отдельных термов. Одним из наиболее популярных алгоритмов этого вида является алгоритм CMU основанный на представлении новости как вектора в пространстве термов [4].

Для представления i-ой новости в этом алгоритме используется вектор, координаты которого соответствуют всем возможным термам, а значения координат вычисляются по метрике TFIDF [3]:

$$v_{ji} = \frac{\left(1+\log_2\left(\mathrm{f_i}\left(c_j\right)\right)\right)\log_2\left(\left|N_U\right|/\mathrm{f_U}\left(c_j\right)\right)}{\sqrt{\sum_{m=1}^{|C|}\left(\left(1+\log_2\left(\mathrm{f_i}\left(c_m\right)\right)\right)\log_2\left(\left|N_U\right|/\mathrm{f_U}\left(c_m\right)\right)\right)^2}}, \qquad (1)$$

где v_{ji} – значение соответствующее j-ому терму в i-ой web-новости;

C – множество уникальных термов из всех web-новостей;

U – множество всех доступных для анализа web-новостей;

$\mathrm{f_i}\left(c_j\right)$ – частота появления терма c_j в i-ой web-новости;

$\mathrm{f_U}\left(c_m\right)$ – частота появления терма c_m во всех доступных web-новостях;

N_U – множество всех доступных для анализа web-новостей.

Для определения тематической близости i-ой новости к новостям посвященным k-ому событию, используют функцию для определения тематической близости документов:

$$\mathrm{sim}(l,i)=\frac{\sum_{j=1}^{|C|} v_{jl}\cdot v_{ji}}{\sqrt{\left(\sum_{j=1}^{|C|} v_{jl}^2\right)\cdot\left(\sum_{j=1}^{|C|} v_{ji}^2\right)}}, \qquad (2)$$

где i и l – задают сравниваемые документы.

Решение о том принадлежит ли i-ая новость k-ому кластеру, принимается на основании формулы 3:

$$\mathrm{SIM}(N_k,i)=\begin{cases}\left(1-|N_m|/|N_d|\right)\mathrm{sim}(\mathrm{cen}(N_k),i), & \text{для } \Delta t\le t_w\\ 0, & \text{для } \Delta t> t_w,\end{cases} \qquad (3)$$

где t_w – некоторый экспериментально установленный интервал времени называемый временным окном (в реализованной системе в качестве такого интервала были выбраны одни сутки);

N_d – задает web-новости сгенерированные в период времени продолжительностью t_w окончившийся в момент появления i-ой новости;

N_m – множество web-новостей выпущенных в период между выходом последней новости освещающей k-ое событие и публикацией i-ой новости;

$\mathrm{cen}\left(N_k\right)$ – центроид множества web-новостей оцененных как освещающие k-ое событие;

Δt –временной интервал между выходом последней web-новости которая была отмечена как посвященная k-ому событию и i-ой новостью [3].

Алгоритм CMU использует формулы 1 – 3 и состоит из следующих шагов:

- создание образа первой web-новости с использованием метрики TFIDF;

- задание первой web-новости в качестве центроида множества web-новостей посвященных некоторому событию (в дальнейшем, центроид множества web-новостей посвященных событию, выбирается при каждом определении новой web-новости как освещающей данное событие);

- в соответствии с формулой 3 и экспериментально выбранным пороговым значением функции SIM, анализируемая web-новость отмечается как посвященная одному или нескольким отдельным событиям;

- если, после предыдущего шага не было найдено событие, освещаемое web-новостью, то ее отмечают как посвященную новому событию [3].

Одним из основных недостатков алгоритма CMU является то, что он хоть и учитывает возможность существования новостей посвященных сразу нескольким событиям, но не предусматривает вариант, при котором освещение события начинается в новости посвященной также и тем событиям, которые уже были освещены. Он присущ также и другим алгоритмам, рассматривающим web-новость как набор термов [3].

Этот недостаток можно исправить с помощью дополнительного анализа текста новости, позволяющего определить присутствие в новости информации, освещающей ранее не освещавшееся событие. Известно, что с течением времени содержание web-новостей освещающих одно и тоже событие изменяется [3]. Можно предположить, что web-новости освещающие событие постепенно меняют свою направленность от предсказания события к констатации, а затем и к обсуждению его последствий. Понятно, что этот процесс всегда проходит в одной и той же последовательности, но не всегда в нем присутствуют все три этапа, например, этап предсказания события может отсутствовать, если событие не предсказуемо. На основании этого можно утверждать что, если все известные события, которые освещает новость, уже констатировались как факты в предыдущих новостях, а анализируемая новость предсказывает некоторое событие, то либо она является устаревшей, либо посвящена неизвестному до сих пор событию.

Для анализа характера новости предположительно можно использовать наличие в тексте новости характерных слов. Так глаголы будущего времени, характерны для предсказывающих новостей, глаголы настоящего времени причастного склонения –

для констатирующих новостей. А глаголы прошедшего времени причастного склонения – для новостей, имеющих направленность на обсуждение.

Теоретически такой подход мог бы дать хорошие результаты, однако на практике проявляется как минимум две трудности его реализации. Первая заключается в том что разного рода события в новостях в основном освещаются либо находясь только на одной фазе развития, например, ЧП освещают только тогда когда оно произошло либо разные стадии освещенности события могут быть оценены как разные события, например, принятие решения о месте и времени проведения олимпиады и проведение олимпиады. Вторая трудность в применении данного подхода к новостям связана с тем, что одно и то же событие в один и тот же момент времени разные люди могут описать как в прошедшем, так и в настоящем времени, например, «президент заявил, что ***» или «*** заявляет президент». Для преодоления этих трудностей необходимы серьезные исследования в области анализа текста.

Возможно, эти исследования будут проведены и использованы при создании следующих версий системы, но данная версия в чистом виде не использует описанные принципы. Тем не менее, элементы описанных принципов были использованы и при создании тестировавшейся версии программы. А именно, использовалось то обстоятельство, что новость, имеющая более одной направленности, содержит в себе информацию о нескольких событиях. То, что каждое из событий освещаемых в новости, возможно, находится на единственной для себя стадии освещенности, ни как не влияет на предложенное предположение. Кроме того, тот факт что одно и тоже событие в одно и тоже время разные люди могут описать по разному, также не принципиален. Т.к. каждую новость пишет только один автор и, следовательно, одно и тоже событие в одной и той же новости везде будет описано с использованием оборотов относящихся только к одной стадии освещенности.

На основании приведенных суждений можно говорить о том, что не все новости одинаково характерны для сюжетов, в которые они входят. Точнее, новости содержащие сведения, относящиеся к разным событиям, в целом не могут быть использованы в качестве характерных образцов для каких-либо отдельных событий, так как содержат примеси других событий. Эту идею можно учесть при вычислении центроидов кластеров посвященных событиям. В соответствии с этой идеей новости, имеющие не одну

направленность не должны учитываться при подсчете центроидов, что может помочь обеспечить большую точность определения свойств новостей характерных для каждого события.

3. Реализация системы

В основу реализованной системы были положены описанные в предыдущем разделе принципы. Точнее, по описанному алгоритму производилось выделение новостей посвященных отдельным событиям. При этом для определения направленности текста новостей (а точнее выявления временных форм глаголов) использовалась разработанная Yandex находящаяся в свободном распространении система mystem. Эта система также использовалась для исключения из текста новости стоп-слов. Такими словами считались все те слова, которые не являлись ни существительными, ни прилагательными, ни глаголами.

Формирование новостных сюжетом было реализовано в два этапа. Первый этап был сделан на основании того принципа, что если была новость, содержащая информацию про два события, то эти события как-то взаимосвязаны и, следовательно, входят в один сюжет. В основу второго этапа была заложена формула 3. Но, в отличии от ее использования для выбора новостей относящихся к одному событию было внесено три коррекции:

- сравнивались не новости с кластерами новостей, а пары кластеров;
- порог схожести был выбран примерно в два раза меньшим (т.е. было смягчено условие схожести);
- временное окно было увеличено до бесконечности (т. е. по сути, использовалась только первая часть формулы).

4. Результаты тестирования

В ходе создания системы в тестовых выборках были обнаружены новости, полностью написанные на английском языке и новости, не содержащие ни одного слова (т.е. с исключительно числовой информацией), такие новости не вошли в эксперимент вообще. Остальные же новости, включая новости, содержащие не полный текст, а только его фрагменты были проанализированы на равных условиях.

В результате работы системы были сформированы события и сюжеты из новостей, опубликованных в течение трех разнесенных во времени недель: обычной недели, недели, во время которой произошла отставка Шеварднадзе, и недели, во время которой произошли взрыв в Ессентуках и выборы в Госдуму. Из-за временного разброса новостей каждое событие было описано новостями из одной недели. При этом характерным было то, что подавляющее большинство событий оказались представленными только одной новостью. С другой стороны были единичные события представленные значительным количеством новостей. Причем рекордные количества новостей посвященных одному событию в каждой неделе отличались значительно. Самый большой «рекорд», 95 новостей посвященных одному событию, был получен при анализе недели, во время которой произошли взрыв в Ессентуках и выборы в Госдуму, а самый маленький 21 новость, посвященная одному событию – при анализе недели, во время которой произошла отставка Шеварднадзе. При этом внутри каждой недели переход освещенности событий был плавный.

Наличие большого количества событий освещенных единичными новостями совпало с предоставленными результатами ручной сортировки новостей. Разницу же между рекордно освещенными событиями проверить не удалось, т.к. ручной анализ был произведен только среди новостей одной недели. Тем не менее, можно предположить, что такой разброс вызван тем, что во время недели с выборами Госдумы были более резонансные события, чем в остальные недели. Причем маленький уровень резонансности недели с отставкой Шеварднадзе, а не обычной недели можно объяснить только искусственным снижением резонансности одних событий и искусственным же поднятием резонансности других событий.

При выделении из новостей сюжетов наблюдалась та же картина малова количества больших сюжетов и подавляющего большинства сюжетов из единичных новостей. Что так же присутствовало и при ручном способе анализа новостей. Однако, если в случае анализ событий можно было говорить о различиях между неделями, то в случае новостных сюжетов о таких вещах говорить не возможно, т.к. из-за снятия контроля за временем выхода новости было получено значительное количество новостных сюжетов содержащих новости из разных недель. В качестве примера такого сюжета можно привести сюжет, в качестве темы которого можно было бы

выбрать заголовок «чемпионат хоккея в России». В этот сюжет системой были правильно занесены одна новость из недели с отставкой Шеварднадзе, и по две новости из остальных двух недель, кроме того, в этот сюжет ошибочно попала новость об изменении курса валют.

В ходе эксперимента были подсчитаны усредненные метрики Precision и Recall, которые характеризовали соответственно точность и полноту работы системы. При этом тот факт, что идеальная система (т.е. человек) при анализе новостей разделила их на не пересекающиеся кластеры, а тестируемая система относила одни и те же новости к разным событиям, был учтен двумя разными способами. Первый способ заключался в отнесении проблемной (посвященной, по мнению системы, нескольким событиям) новости к одному произвольно выбранному событию. Вторым же способом было объединение событий распознанных системой, в том случае если они освещались одной и той же новостью. Таким образом, было проведено два замера характеристик системы, причем характеристики были рассчитаны не на работу системы в целом, а отдельно на качество распознание системой событий, сюжетов и надсюжетов. В таблице 1 приведены результаты тестирования системы.

Таблица 1 – Метрики качественности системы

Тип кластеров создаваемых системой	С применением случайного выбора кластера		С применением объединения кластеров	
	Точность	Полнота	Точность	Полнота
Событие	0.709	0.598	0.705	0.606
Сюжет	0.293	0.472	0.293	0.472
Надсюжет	0.577	0.098	0.577	0.098

Полученные характеристики системы при выделении новостей относящихся к одному событию сравнимы с характеристиками других принимавших участие в эксперименте системами, что свидетельствует о жизнеспособности ее основных принципов. А вот качество выбора новостей относящихся к одному сюжету было оценено значительно хуже, чем у других систем, что свидетельствует о необходимости пересмотра алгоритмов выделения сюжетов.

При анализе качества формирования системой кластеров новостей освещающих одно событие, можно заметить три особенности:

- как точность, так и полнота результатов системы слабо зависят от метода учета новостей отнесенных системой к нескольким событиям, это свидетельствует о том, что такие новости существовали, но либо их было мало, либо методы их учета схожи между собой;
- в обоих случаях точность лучше полноты, что может быть вызвано слишком сильным разбиением новостей, т.е. слишком малым временным окном или слишком большим порогом схожести при применении формулы 3;
- точность и полнота, отличаются не сильно и далеки от идеальных значений, что свидетельствует о том, что описанный недостаток является не основным недостатком системы.

При анализе качества формирования системой кластеров новостей относящихся к одному сюжету разницы между способами оценки новостей попавших сразу в несколько кластеров вообще нет. Это вызвано тем, что при формировании этих кластеров система объединяла все те группы новостей посвященные событиям которые содержали одну и туже новость, что привело к тому что каждая новость попала только в один сюжет. Вмести с тем, следует заметить, что в случае анализа новостных сюжетов, лучшее значение имеет полнота, это дает возможность предполагать, что система относила слишком большое количество событий в один сюжет. С другой стороны, как и в случае с анализом событий, даже более хорошая характеристика системы при анализе сюжетов далека от идеальной. На основании этого можно заключить, что тут так же присутствует еще какой-то негативный фактор, причем этот фактор связан не только с не корректным формированием событийных кластеров, т.к. характеристики системы при создании сюжетов хуже, чем при выделении событий.

При анализе качественности распознавания системой надсюжетов выделенных при анализе работы системы заметно, что точность значительно превышает полноту ответов системы, но говорить о недостатках системы при выделении надсюжетов сложно т.к. непосредственно система не формировала надсюжеты. Зато можно заметить, что характеристики системы при формировании, как сюжетов, так и надсюжетов хуже, чем ее характеристики при выделении событий. Это можно объяснить тем, что при

формировании событий погрешность была вызвана в основном неправильным отнесением новостей в кластеры событий далеких по смыслу от этих новостей.

На основании приведенных наблюдений можно предположить, что одним из основных недостатков системы является не достаточно точная работа со стоп-словами, т.к. это одна из наиболее очевидных причин, по которой новости разной тематики могут быть оценены системой как подобные. Примером такой ошибки системы может служить упоминавшийся сюжет о чемпионате хоккея в России. А о массовости подобных ошибок можно судить по приведенной таблице 1.

5. Выводы

Тестировавшаяся система с одной стороны базируется на давно используемом алгоритме CMU, а с другой содержит элементы анализа смысла текста характерные для современных алгоритмов работающих с содержимым Интернета. Из-за того, что элементы анализа смысла текста используемые в алгоритме сравнительно примитивные, его применение к тяжело формализуемому русскому языку вполне реально.

Недостатком системы является недостаточная степень проработанности вспомогательных алгоритмов (таких как удаление стоп-слов), а так же не достаточная степень настройки параметров основного алгоритма. Кроме того, в тестировавшейся версии системы не были реализованы все заложенные в ее основу идеи.

Не смотря на имеющиеся у системы недостатки можно говорить о ее жизнеспособности и целесообразности дальнейших ее модификаций.

Литература:

1. Д.В. Ландэ, Основы интеграции информационных потоков, Монография. – К.: Инжиниринг, 2006. – 240 с.

2. Д.В. Ландэ и др., Основы моделирования и оценки электронных информационных потоков, Монография. – К.: Инжиниринг, 2006. – 176 с.

3. http://www.cs.bilkent.edu.tr/tech-reports/2001/BU-CE-0110.ps.gz

4. http://www.i-teco.ru/article104.html

5. http://ru.wikipedia.org/wiki/

Modernization of calculation centroid in algorithm CMU

© Suyargulova E. B.
Kharkiv national university of radioelectroniks
s_yv_b@mail.ru

© Vechur A. V.
Kharkiv national university of radioelectroniks
vechur@yahoo.com

Abstract

The paper contains the analysis of program system intended for allocation from the general stream of web-news of clusters of news concerning to one event, and as to one news scene. In paper the basic ideas put in a basis of system, the general algorithm of work of system, and as results of its testing are resulted.