Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии

А.В. Коршунов
korshunov@ispras.ru

Аннотация. В статье описывается способ извлечения ключевых терминов из сообщений микроблогов с использованием информации, полученной путём анализа структуры и содержимого интернет-энциклопедии Википедия. Работа алгоритма основана на расчёте для каждого термина его "информативности", т.е. оценки вероятности того, что он может быть выбран ключевым в тексте. В ходе тестирования разработанный алгоритм показал удовлетворительные результаты в условиях поставленной задачи, существенно опережая аналоги. В качестве демонстрации возможного применения разработанного алгоритма был реализован прототип системы контекстной рекламы. Сформулированы также варианты использования информации, полученной путём анализа сообщений Twitter, для реализации различных вспомогательных сервисов.

Ключевые слова: Информационный поиск; извлечение ключевых терминов; обработка естественного языка; анализ текста; семантический анализ; микроблоггинг; Твиттер; Википедия; контекстно-зависимая реклама.

1. Введение

На сегодняшний день одной из самых важных и заметных областей Web 2.0, ключевым принципом которой является участие пользователей в работе сайтов, являются сетевые дневники, или веб-логи, сокращённо называемые *блогами*. Концептуальным развитием блогов, обусловленным их широкой социализацией, являются микроблоги, которые имеют ряд характерных особенностей: ограниченная длина сообщений, большая частота публикаций, разнообразная тематика, различные пути доставки сообщений и т.д.

Первый и наиболее известный сервис микроблогов *Twitter* был запущен в октябре 2006 г. компанией *Obvious* из Сан-Франциско. К настоящему времени постоянно растущая аудитория сервиса составляет десятки миллионов человек. Очевидно, что автоматизированное выделение наиболее значимых терминов из потока сообщений, генерируемого сообществом Twitter, имеет практическое значение как для определения интересов различных групп

пользователей, так и для построения индивидуального профиля каждого из них.

Однако нужно отметить, что классические статистические методы экстракции ключевых терминов, основанные на анализе коллекций документов, малоэффективны в данном случае. Это обусловлено чрезвычайно малой длиной сообщений (до 140 символов), их разнообразной тематикой и отсутствием логической связи между собой, а также обилием редко используемых аббревиатур, сокращений и элементов специфического микросинтаксиса.

Для решения этой проблемы в представленной работе относительная значимость терминов в анализируемом контексте определяется с помощью данных о частоте их использования в качестве ключевых в интернет-энциклопедии Википедия. Работа алгоритма основана на расчёте для "информативности" каждого термина, т.е. оценки вероятности того, что он может быть выбран ключевым в тексте. В дальнейшем к анализируемому набору терминов применяется ряд эвристик, результатом которых является список терминов, сочтённых ключевыми.

2. Обзор Twitter

Возникновение и последующее развитие идеи создания сервиса *микроблогов* современные исследователи Интернета вполне обоснованно считают результатом процесса интеграции концепции *социальных сетей* с сетевыми дневниками – *блогами* [1].

По определению, данному Walker в [2], *блогами* называются часто обновляемые веб-сайты, состоящие из содержащих различную информацию записей, размещённых в обратном хронологическом порядке.

Характерные черты *блоггинга* могут быть описаны с использованием 3 ключевых принципов [3]:

- содержимое блогов представляет собой короткие сообщения;
- сообщения имеют общее авторство и находятся под контролем автора;
- возможна агрегация множества потоков сообщений от разных авторов для удобного чтения.

Эти принципы применимы также и для *микроблоггинга* [4]. Однако в то время как для блоггинга публикация и агрегация сообщений являются задачами для разных программных продуктов, сервисы микроблоггинга предоставляют все перечисленные возможности сразу.

Кроме того, микроблоггинг учитывает потребность пользователей в более быстром режиме коммуникации, чем при обычным блоггинге. Поощряя более короткие сообщения, он уменьшает время и мысленную работу, необходимые для создания контента. Это также является одним из главных отличительных

от блоггинга признаков. Другое отличие заключается в частоте обновлений. Автор обычного блога обновляет его, в среднем, раз в несколько дней, в то время как автор микроблога может обновлять его несколько раз в день.

Основные функции наиболее популярного на сегодняшний день сервиса микроблогов Twitter представляют собой очень простую модель. Пользователи могут отправлять короткие сообщения, или *твиты*, длиной не больше 140 символов. Сообщения отображаются как *поток* на странице пользователя. В терминах социальных связей, Twitter даёт возможность пользователям *следовать* (*follow*) за любым числом других пользователей, называемых *друзьями*. Сеть контактов Twitter асимметрична, т.е. если один пользователь следует за другим, то второй не обязан следовать за ним. Пользователи, следующие за другим пользователем, называются его *последователями* (*followers*).

Пользователи имеют возможность указать, желают ли они, чтобы их твиты были доступны *публично* (появляются в обратном хронологическом порядке на главной странице сервиса и на странице самого пользователя, называемой *микроблогом*) или *приватно* (только последователи пользователя могут видеть его сообщения). По умолчанию, все сообщения доступны любому пользователю.

Для облегчения понимания внутреннего устройства Twitter целесообразно ввести 2 концепции: *элемент* как отдельное сообщение и *канал* как поток элементов, чаще всего принадлежащих одному пользователю [3]. Способы взаимодействия концепций между собой представлены в табл. 1.

	Канал	**Элемент**
Канал	следование	ретвит
Элемент	@-ссылка	ответ

Табл. 1. Способы взаимодействия концепций Twitter между собой.

Ниже приведено краткое описание каждого из способов взаимодействия:

- *следование* — один канал имеет другой в своей сети и читает его обновления;
- *@-ссылка* — текст элемента может ссылаться на другой канал с помощью конструкции @<имя_канала>;
- *ретвит* — пользователи берут элементы из чужих каналов и помещают в свои с добавлением @-ссылки на источник и, в некоторых случаях, своего комментария;
- *ответ* — один элемент является прямой реакцией на предыдущий.

В заключение общего описания особенностей Twitter можно отметить, что необходимым условием для успешного принятия пользователями новой технологии (или нового способа применения существующего инструмента) является позитивное отношение к её потенциалу. Компания Gartner добавила микроблоггинг в свой *"цикл ажиотажа"* (*hype cycle*) в 2008 году [5], предсказывая резкий рост популярности этого явления. Согласно Gartner, передовые компании исследуют потенциал микроблоггинга, чтобы усовершенствовать другие социальные информационные средства и каналы. Всё это говорит о том, что микроблоги являются одним из самых перспективных и динамично развивающихся сегментов Интернета.

Вместе с тем, микроблоги являются всё ещё относительно новым явлением среди онлайновых социальных сетей и на данном этапе недостаточно исследованы.

2.1. Особенности сообщений

Помимо наличия «ретвитов» и @-ссылок, в «твитах» могут также присутствовать и другие элементы специфического *микросинтаксиса*, задачей которого является представить часто используемые понятия в сокращённой общеупотребительной форме, а также расширить набор инструментов для повышения информативности сообщений в условиях ограниченного их размера.

Основной частью микросинтаксиса являются *слэштеги*, каждый из которых состоит из символа «/» и *указателя*, который и определяет смысл слэштега. Эти элементы используют для различных целей. Например, «*/via*» - для ссылки на автора сообщения при «ретвите»; «*/by*» - для ссылки на автора исходного сообщения, если оно является результатом цепочки «ретвитов»; «*/cc*» – для указания тех подписчиков микроблога, которым в первую очередь адресовано сообщение и т.д.

Использование слэштегов полностью является инициативой самих пользователей, поэтому не существует единых правил по их использованию. Приведённое описание отражает наиболее популярные способы применения слэштегов к настоящему моменту. Существуют, однако и другие рекомендации, которые тоже заслуживают внимания, т.к. каждый пользователь применяет элементы микросинтаксиса по своему усмотрению. Например, возможно объединение всех использованных в сообщении слэштегов без символа «/» в одну группу, ограниченную скобками. Некоторые пользователи размещают слэштеги строго в конце сообщения, предваряя лишь первый из них символом «/» с целью экономии символов.

Поскольку в Twitter не существует простого и удобного способа для группирования «твитов» разных пользователей по тематике, сообщество пользователей пришло к собственному решению: использование *хэштегов*. Они похожи на другие примеры использования тегов (например, для

аннотирования записей в обычных блогах) и позволяют добавить «твиты» в какую-либо категорию.

Хэштеги начинаются с символа «#», за которым следует любое сочетание разрешённых в Twitter непробельных символов; чаще всего это слова или фразы, в которых первая буква каждого слова приведена к верхнему регистру. Они могут встречаться в любой части «твита», зачастую пользователи просто добавляют символ «#» перед каким-либо словом. Другим вариантом использования является добавление популярных хэштегов, таких, как «*#haiku*». При добавлении в сообщение хэштега оно будет отображаться при поиске в потоке сообщений Twitter по этому хэштегу.

К неофициальным, но общепринятым правилам использования хэштегов относится выбор в качестве них терминов, релевантных теме сообщения, а также добавление лишь небольшого количества их в одно сообщение. Это позволяет рассматривать их в качестве терминов, которые с достаточной степенью вероятности отражают общую тематику сообщения.

3. Существующие подходы к извлечению ключевых терминов

Одной из задач извлечения информации из текста является выделение ключевых терминов, с определённой степенью достоверности отражающих тематическую направленность документа. Автоматическое *извлечение ключевых терминов* можно определить как автоматическое выделение важных тематических терминов в документе. Оно является одной из подзадач более общей задачи – автоматической *генерации ключевых терминов*, для которой выделенные ключевые термины не обязательно должны присутствовать в данном документе [6]. В последние годы было создано множество подходов, позволяющих проводить анализ наборов документов различного размера и извлекать ключевые термины, состоящие из одного, двух и более слов.

Важнейшим этапом извлечения ключевых терминов является расчёт их *весов* в анализируемом документе, что позволяет оценить их значимость относительно друг друга в данном контексте. Для решения этой задачи существует множество подходов, которые условно делятся на 2 группы: требующие обучения и не требующие обучения. Под *обучением* подразумевается необходимость предварительной обработки исходного корпуса текстов с целью извлечения информации о частоте встречаемости терминов во всём корпусе. Другими словами, для определения значимости термина в данном документе необходимо сначала проанализировать всю коллекцию документов, к которой он принадлежит. Альтернативным подходом является использование лингвистических онтологий, которые являются более или менее приближёнными моделями существующего набора слов заданного языка. На базе обоих подходов были созданы системы для автоматической экстракции ключевых терминов, однако в этом направлении

постоянно ведутся работы с целью повышения точности и полноты результатов, а также с целью использования методов извлечения информации из текста для решения новых задач [7-13].

Самыми распространёнными схемами для расчёта весов терминов являются *TF-IDF* и различные его варианты, а также некоторые другие (*ATC*, *Okapi*, *LTU*). Однако общей особенностью этих схем является то, что они требуют наличия информации, полученной из всей коллекции документов. Другими словами, если метод, основанный на TF-IDF, используется для создания представления о документе, то поступление нового документа в коллекцию потребует пересчёта весов терминов во всех документах. Следовательно, любые приложения, основанные на значениях весов терминов в документе, также будут затронуты. Это в значительной мере препятствует использованию методов извлечения ключевых терминов, требующих обучения, в системах, где динамические потоки данных должны обрабатываться в реальном времени, например, для обработки сообщений микроблогов [14].

Для решения этой проблемы было предложено несколько подходов, таких как алгоритм *TF-ICF* [15]. В качестве развития этой идеи Mihalcea и Csomai в 2007 году предложили использовать в качестве обучающего тезауруса Википедию [16]. Они применили для расчётов информацию, содержащуюся в аннотированных статьях энциклопедии с вручную выделенными ключевыми терминами. Для оценки вероятности того, что термин будет выбран ключевым в новом документе, используется формула:

$$P(\textit{ключевой термин} \mid W) \approx \frac{\textit{число}(D_{\textit{ключевой}})}{\textit{число}(D_W)}, \qquad (1)$$

где W - термин;

$D_{\textit{ключевой}}$ - документ, в котором термин был выбран ключевым;

D_W - документ, в котором термин появился хотя бы один раз.

Эта оценка была названа авторами *keyphraseness* (в данной работе определена как *информативность*). Она может быть интерпретирована следующим образом: «чем чаще термин был выбран ключевым из числа его общего количества появлений, тем с большей вероятностью он будет выбран таковым снова».

Информативность может принимать значения от 0 до 1. Чем она выше, тем выше значимость термина в анализируемом контексте. Например, для термина «*Of course*» в Википедии существует только одна статья, посвящённая песне американского исполнителя, поэтому он редко выбирается ключевым, хотя встречается в тексте очень часто. Значение его информативности, таким образом, будет близко к 0. Напротив, термин «*Microsoft*» в тексте любой

статьи почти всегда будет выделен ключевым, что приближает его информативность к 1.

Данный подход является довольно точным, т.к. все статьи в Википедии вручную аннотируются ключевыми терминами, поэтому предложенная оценка их реальной информативности является лишь результатом обработки мнений людей.

Вместе с тем, эта оценка может быть ненадёжной в тех случаях, когда используемые для расчётов значения слишком малы. Для решения этой проблемы авторы рекомендуют рассматривать только те термины, которые появляются в Википедии не менее 5 раз.

В заключение обзора методов извлечения ключевых терминов нужно сказать, что для расчёта веса термина w_i в данной работе использовалась формула:

$$w_i = TF_i \cdot K_i \qquad (2)$$

,

где i - порядковый номер термина;

TF_i - частота термина в анализируемом сообщении;

K_i - информативность термина по данным Википедии.

TF означает *частоту термина* (*Term Frequency*). Значение этого компонента формулы равно отношению числа вхождения некоторого термина к общему количеству терминов сообщения. Таким образом, оценивается важность термина t_i в пределах отдельного сообщения.

4. Описание алгоритма

4.1. Извлечение информации из Википедии

Одним из важнейших этапов разработки системы явилась обработка XML-дампа всех статей английской Википедии по состоянию на июль 2009 г. Целью анализа был расчёт *информативности* для всех терминов Википедии по формуле (1).

Нужно отметить, что для одной *концепции* в словаре Википедии может быть несколько *синонимов*. Например, термин «*IBM*» имеет несколько синонимов: «*International Business Machines*», «*Big Blue*» и т.д. Так как в разработанной системе отсутствует этап *разрешения лексической многозначности* терминов, то было недопустимо, чтобы синонимы имели различные значения информативности. Поэтому было принято считать, что информативность всех синонимов одной концепции становится одинаковой, исходя из общей статистики для всех них.

Кроме того, согласно рекомендациям авторов методики расчёта информативности [16], были исключены термины, которые были найдены *менее чем в 5 статьях*. Если пропустить этот шаг, то результирующее значение зачастую становится недостоверным и не позволяет корректно оценить относительную значимость термина в контексте. В результате данного этапа БД содержит 5 445 377 терминов с рассчитанной для них информативностью.

4.2. Извлечение ключевых терминов

Общая архитектура разработанной системы представлена на рис. 1.

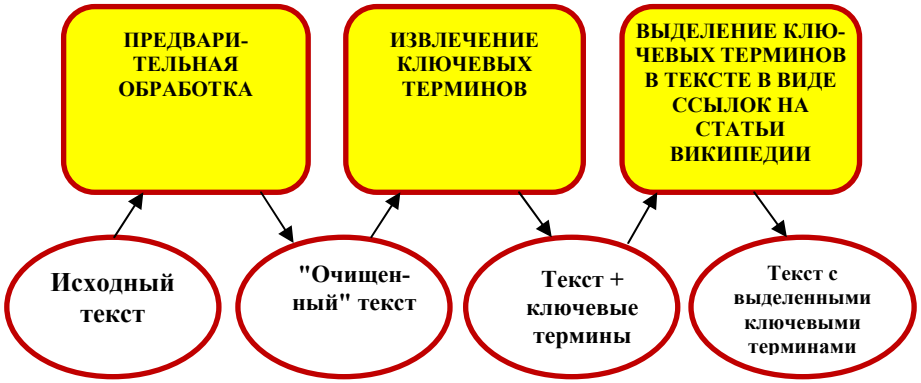

Рис. 1. Общая архитектура системы.

Для получения информации о сообщениях пользователя с сервера Twitter использовался Perl-модуль *Net::Twitter*. Результатом взаимодействия с Twitter API является получение некоторого числа последних сообщений его аккаунта, иначе называемых *timeline*.

Для решения поставленной задачи был выбран метод *statuses/friends_timeline*, возвращающий сообщения друзей пользователя. Для тестирования были созданы аккаунты, подписанные на обновления статуса единственного друга. При этом в самом аккаунте не публиковалось никаких сообщений. Таким образом, результат вызова данного метода содержит лишь необходимое количество сообщений одного пользователя Twitter, что и требуется в качестве исходных данных.

В процессе *предварительной обработки* текста, или *препроцессинга*, содержимое полученных с сервера Twitter сообщений преобразуется к формату входных данных для алгоритма извлечения ключевых терминов.

Помимо стандартных для этого этапа операций, производится также удаление *слэштегов* и *@-ссылок*, а также служебных слов «*RT*» и «*RETWEET*», то есть тех элементов специфического синтаксиса Twitter, которые не несут

смысловой нагрузки и являются *стоп-словами* в терминах обработки естественного языка. На этом этапе также производится извлечение *хэштегов*, которые в последующем обрабатываются наравне с остальными словами.

Очевидно, однако, что для составления списка кандидатов в ключевые термины недостаточно лишь исходной последовательности слов. Исходя из небольшой длины сообщений, представляется возможным сделать алгоритм максимально избыточным с целью повышения *полноты* результатов, то есть не пропустить ни один возможный ключевой термин, из скольких бы слов он ни состоял. С этой целью производится поиск *всех* возможных *N-грамм*, то есть последовательностей идущих друг за другом слов. Количество N-грамм Q_k, которые можно получить из k слов (включая N-граммы из одного слова), таким образом, равно:

$$Q_k = \sum_{i=1}^{k} i \qquad (3)$$

Все полученные на этом этапе термины добавляются в массив возможных ключевых терминов.

Завершающим этапом препроцессинга является *стоплистинг*, т.е. удаление из полученного массива тех слов, которые не несут существенной смысловой нагрузки. Важно отметить, что стоплистинг выполняется лишь после извлечения N-грамм. Таким образом, стоп-слова могут входить в составные термины, но удаляются из списка кандидатов, если встречаются в нём сами по себе. На этом этапе используется стоп-лист системы *SMART* [17].

На этапе расчёта весов терминов-кандидатов для каждого из них запрашивается значение *информативности* из базы данных. Вторым необходимым для расчётов показателем является частота встречаемости термина *TF*. Для каждого найденного в БД термина его вес рассчитывается по формуле (2).

Общий принцип извлечения ключевых терминов заключается в анализе заданного числа сообщений и определении *порогового значения веса* для каждого из них. Те термины, веса которых *больше или равны* пороговому значению, считаются ключевыми.

Изначально пороговым считается *среднее арифметическое* значение веса для всех терминов-кандидатов. Все последующие операции призваны уточнить его и улучшить результаты работы алгоритма в целом.

Следующим этапом является обработка массива *хэштегов*, полученного на этапе препроцессинга. Предполагается, что с их помощью пользователь явно указывает термины, определяющие тематику сообщения. Поэтому логично предположить, что пороговое значение для всего сообщения не должно быть выше минимального веса среди его хэштегов. Основываясь на этом предположении, вес каждого из хэштегов (если он был найден в БД)

сравнивается с текущим пороговым значением и понижает его в случае, если оно больше.

Если после обработки хэштегов пороговое значение осталось равным 0 (в случае, когда они не были указаны или когда ни один из них не был найден в БД) либо превышает найденное среднее значение, оно принимается равным среднему. Такая ситуация на практике встречается чаще всего, так как пользователи редко явно указывают тематические термины. Однако в противном случае такой подход существенно улучшает результаты работы.

Нужно отметить, что сообщения обрабатываются в порядке, обратном поступлению с сервера, то есть в прямом хронологическом. Такой подход представляется логичным и учитывает специфику сервиса блогов в целом: пользователь может написать сообщение на какую-то тему, а затем вернуться к ней снова. Однако во втором сообщении, помимо тех терминов, которые были выбраны ключевыми в первом, могут быть другие, более информативные термины, за счёт которых пороговое значение для второго сообщения будет завышено, и ключевые термины из первого сообщения не будут выделены. Чтобы избежать этого, в системе имеется отдельный массив, содержащий все ранее извлечённые ключевые термины. Тогда при обработке очередного сообщения термины из этого массива будут безусловно извлечены и снизят пороговое значение веса для данного сообщения.

В этом контексте важно, что при непосредственном выборе ключевых терминов кандидаты обрабатываются в порядке *возрастания* их весов. Таким образом, если вес какого-либо из кандидатов ниже порогового, но он выбирается ключевым из-за того, что присутствует в массиве ранее извлечённых терминов, то пороговое значение становится равным его весу, и все следующие термины автоматически попадают в список ключевых.

Результатом работы алгоритма является список отсортированных в порядке *убывания* весов ключевых терминов.

5. Взаимодействие с Amazon API

В качестве демонстрации возможного применения разработанного алгоритма было реализовано получение с сервера интернет-магазина *Amazon* описаний товаров, *релевантных* найденным ключевым терминам.

Для взаимодействия с Amazon REST API используется Perl-модуль *Net::Amazon*, который предоставляет удобный доступ к большинству функций этого программного интерфейса. При этом осуществляется поиск по всем товарам интернет-магазина, в названии или описании которых встречается искомый термин. Для необходимого количества наиболее подходящих товаров выводятся название, цена, год издания и изображение. Название товара представляет собой ссылку на его страницу на сайте Amazon.

При подключении к серверу используется *секретный ключ разработчика*, который предоставляется при регистрации в сервисе Amazon и позволяет

вести статистику переходов на страницы товаров с сайтов сторонних разработчиков. Если после перехода по ссылке пользователь приобретёт товар, то, согласно *партнёрской программе* Amazon, владелец сайта может получить некоторое вознаграждение, равное части стоимости товара. Подобная схема может быть реализована не только для Amazon, но также и для любого другого интернет-магазина, имеющего партнёрскую программу.

6. Результаты экспериментов

Результатом работы системы является HTML-страница, разбитая на блоки, каждый из которых соответствует одному сообщению. В блоке выводится текст оригинального сообщения с указанием его автора, затем – текст после препроцессинга и, наконец, тот же текст после обработки. В тексте сообщения на выходе работы системы найденные ключевые термины являются ссылками на соответствующие статьи Википедии.

Для всех найденных ключевых терминов строится таблица, каждая строка которой содержит термин, его вес и найденные релевантные товары из интернет-магазина. Ниже выводятся среднее и пороговое значения веса. Последней частью выходных данных является список терминов, которые были найдены в базе, но не были отнесены к ключевым.

Эффективность алгоритмов извлечения ключевых терминов обычно оценивается путём сравнения результатов их работы с ключевыми терминами, извлечёнными *вручную*. Критерии качества работы основаны на числе соответствий между фразами, выбранными алгоритмом и человеком [6].

Для тестирования работы системы было создано несколько тестовых аккаунтов, каждый из которых был «подписан» на обновления статусов различных известных в IT-сообществе пользователей Twitter. В качестве основного аккаунта для тестирования был выбран *semtweettest2*, который был «подписан» на обновления блога Tim O'Reilly (*timoreilly*), книгоиздателя и общественного деятеля, который имеет свыше 1 400 000 подписчиков. Сообщения в этом блоге отличаются чрезвычайно разнообразной тематикой, в них часто используются различные именованные сущности (имена людей, названия компаний и мероприятий, географические названия), которые представляют реальный интерес в настоящий момент. Кроме того, автор блога полностью использует возможности микросинтаксиса Twitter. Всё это в совокупности даёт основания полагать, что результаты работы разработанной системы на сообщениях блога *timoreilly* позволяют достоверно оценить эффективность алгоритма.

Для сравнения результатов работы алгоритма с существующими аналогами была выбрана система *Alchemy API* [18], которая предоставляет демонстрационный доступ к своим функциям в онлайн-режиме. В качестве исходных данных эта система использует текстовый документ, а возвращает этот же документ с выделенными ключевыми терминами.

Каждое из выбранных для тестирования сообщений было проанализировано с помощью реализованной системы и демонстрационной версии Alchemy API. Так как нет необходимости рассчитывать точность, полноту и F-меру для каждого из них, то весь массив сообщений был принят за один документ, из которого извлекались ключевые термины.

Всего из выбранных 50 сообщений вручную было выделено 180 ключевых терминов, 28 из которых являются частями других, более длинных составных терминов. Максимальная длина выделенного вручную термина равна 3 словам. Максимальная длина термина, выделенного системой, равна 6 словам.

Результаты тестирования приведены в табл. 2.

Метод	Точность, %	Полнота, %	F-мера, %
Alchemy API	18,9	43,6	26,4
Разработанная система	40,0	68,6	50,5

Табл. 2. Результаты тестирования работы системы

По результатам тестирования можно сделать вывод, что разработанная система достаточно эффективно функционирует в условиях поставленной задачи. Кроме того, по качеству результатов она превосходит выбранную для сравнения систему Alchemy API.

Одной из возможных причин снижения качества результатов является большое количество *именованных сущностей* в текстах обработанных сообщений, причём большинство из них указывают на людей, события или компании, ставшие популярными лишь недавно. Очевидно, что, так как БД терминов из Википедии соответствует её состоянию на июль 2009 года, то многие из актуальных в настоящий момент именных сущностей отсутствуют в ней (например: «*CityCamp*», «*CrisisCamps*»). Другой причиной служит частое использование *сокращений*, многие из которых не являются общепринятыми, например: «*webops*» для «*web operations*», «*gov't*» для «*government*», «*gov20*» и «*Gov 2.0*» для «*Government 2.0*».

7. Заключение

В ходе выполнения работы был разработан алгоритм извлечения ключевых терминов из минимально структурированных текстов сообщений микроблогов. Проведённое экспериментальное исследование показало, что алгоритм работает корректно и эффективно. В качестве примера возможного практического использования результатов в рамках разработанной системы реализован прототип системы контекстной рекламы с отображением описаний релевантных товаров из интернет-магазина.

В продолжение начатой работы планируется создание сервиса, основанного на механизме аутентификации *OAuth*, который используется в большинстве современных приложений к Twitter. Пользователь при этом предоставляет доступ к данным своего аккаунта, в том числе к текстам всех сообщений. Одной из возможных перспектив использования разработанного алгоритма является возможность указания пользователем нескольких ключевых терминов с целью просмотра только тех сообщений, которые их содержат. Возможна фильтрация не только по извлечённым с помощью Википедии терминам, но и путём простого поиска по текстам сообщений после препроцессинга. Такой сервис был бы востребованным [19] и позволял бы привлечь целевую аудиторию для показа контекстной рекламы.

Литература

[1] Martin Ebner. *Microblogging - more than fun?* - Proceedings of IADIS Mobile Learning Conference 2008, Inmaculada Arnedillo Sánchez and Pedro Isaías ed., Portugal, 2008, pp. 155-159.

[2] Herman David, Janh Manfred, Ryan Marie-Laure. (éd.), *The Routledge Encyclopedia of Narrative Theory*. London, Routledge, 2005.

[3] Böhringer, M. *Really Social Syndication: A Conceptual View on Microblogging*. - Sprouts: Working Papers on Information Systems, 9(31), 2009.

[4] D.R. Karger, D. Quan (2005). *What would it mean to blog on the semantic web?* - Web Semantics: Science, Services and Agents on the World Wide Web, Selected Papers from the International Semantic Web Conference, Hiroshima, Japan, 07-11 November 2004, 3 (2-3), 2005, 147-157.

[5] *Gartner Highlights 27 Technologies in the 2008 Hype Cycle for Emerging Technologies*. - http://www.gartner.com/it/page.jsp?id=739613, 2008.

[6] P. Turney. *Learning to extract keyphrases from text*. Technical report, National Research Council, Institute for Informational Technology, 1999.

[7] D. Turdakov. *Word sense disambiguation methods*. Programming and Computer Software, 2010, Vol. 36, No. 6, pp. 309-326.

[8] D. Turdakov, S. Kuznetsov. *Automatic word sense disambiguation based on document networks*. Programming and Computer Software, 2010, Vol. 36, No. 1, pp. 11–18.

[9] Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, Denis Turdakov. *Accuracy estimate and optimization techniques for SimRank computation*. - The International Journal on Very Large Data Bases archive. Volume 19 Issue 1, February 2010.

[10] Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, Denis Turdakov. *Accuracy Estimate and Optimization Techniques for SimRank Computation*. - Proceedings of the VLDB Endowment. Volume 1 Issue 1, August 2008.

[11] Maria Grineva, Maxim Grinev, Dmitry Lizorkin. *Effective Extraction of Thematically Grouped Key Terms From Text*. - Proc. of the AAAI 2009 Spring Symposium on Social Semantic Web. - pp. 39-44.

[12] D. Turdakov, D. Lizorkin. *HMM Expanded to Multiple Interleaved Chains as a Model for Word Sense Disambiguation*. - PACLIC 2009: The 23rd Pacific Asia Conference on Language, Information and Computations. - pp. 549-559.

[13] M. Grineva, M. Grinev, D. Lizorkin. *Extracting Key Terms From Noisy and Multitheme Documents*. - WWW2009: 18th International World Wide Web Conference.

[14] M. Grineva, M. Grinev, Alexander Boldakov, Leonid Novak, Andrey Syssoev, D. Lizorkin. *Sifting Micro-blogging Stream for Events of User Interest*. - Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.

[15] Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump, Mark T. Elmore, Ali R. Hurson. *TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams*. - Proc. Machine Learning and Applications, 2006, ICMLA '06, pp. 258-263.

[16] Mihalcea, R., and Csomai, A. 2007. *Wikify!: linking documents to encyclopedic knowledge*. - Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 233-242. New York, NY, USA: ACM.

[17] Salton, G. (1971). *The SMART Retrieval System - experiments in automatic document processing*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

[18] *Alchemy API – Demo*. - http://www.alchemyapi.com/api/demo.html.

[19] Zhao, Dejin and Mary Rosson. *How and why people Twitter: the role that micro-blogging plays in informal communication at work*. - Proceedings of the ACM 2009 international conference on Supporting group work, 2009.

[20] McFedries, P. *All A-Twitter*. IEEE Spectrum, October 2007, 84.

[21] Java, A., Song, X., Finin, T., Tseng, B. *Why we twitter: understanding microblogging usage and communities*. - Proc. WebKDD/SNA-KDD '07, ACM Press (2007).

[22] Krishnamurthy, B., Gill, P., and Arlitt, M. *A few chirps about twitter*. - Proc. WOSP '08. ACM Press (2008).

[23] Honeycutt, C., Herring, S. *Beyond microblogging: Conversation and collaboration via Twitter*. - Proc. HICSS '09. IEEE Press (2009).

[24] Naaman, M., Boase, J., Lai, C.-H. *Is it really about me? Message content in social awareness streams*. - Proc. CSCW 2010, February 6-10, 2010, Savannah, Georgia, USA.

[25] Huberman, B., Romero, D., Wu, F. *Social networks that matter: Twitter under the microscope*. First Monday [Online] 14, 1 (2008).