Анализ методов кластеризации новостного потока

© Кондратьев Михаил Е.

Санкт-Петербургский Государственный Университет
Mikhail.Kondratyev@sun.com

Аннотация

В работе анализируется ряд алгоритмов кластеризации новостной коллекции и приводится описание экспериментов, направленных на выяснение зависимости эффективности и стабильности кластеризации новостного потока от граничных значений алгоритмов.

1 Введение

С развитием сети Интернет все большую популярность завоевывают сетевые новостные агентства, во многом заменившие традиционные средства информации. Большое количество источников информации, резко возросший объем новостных данных и необходимость их быстрой обработки вызвали потребность в создании систем автоматизированного анализа новостного потока. Наиболее известной конференцией, посвященной методам автоматизированной обработки новостных данных, является Topic Detection and Tracking (TDT), где выделяются следующие направления исследований: разбиение потока на сюжеты, идентификация новых событий, определение связей между новостными историями, отслеживание интересующей пользователя информации.

Задачи, связанные с обработкой новостного потока, имеют специфичные особенности, определяемые природой новостных данных. Новостные сообщения являются откликами на события реального мира и поэтому кроме текстового наполнения сюжет объединяют причинно-следственные, временные и прочие факторы. Эта особенность используется некоторыми системами, которые учитывают информацию о времени публикации сообщений [11] и основных действующих лицах [6] при решении задач обработки новостей. Тот факт, что новостные сообщения становятся доступны пользователю в виде потока данных, делает невозможным разбиение уже сформированных сюжетов. Отметим так же, что число сюжетов заранее неизвестно и потенциально неограниченно.

Одной из важнейших задач автоматизированной обработки новостного потока является разбиение новостных сообщений на сюжеты (кластеризация новостного потока). Вследствие потоковой природы новостных данных в значительной части работ используется алгоритм инкрементальной кластеризации [18]:

- выбирается мера близости нового сообщения и кластера;

- для каждого нового сообщения выбирается кластер, наиболее близкий к сообщению;

- в случае, если значение меры близости превышает некоторое пороговое значение, сообщение добавляется в уже существующий кластер;

- в случае если значение меры близости не превысило пороговое значение, создается новый кластер на основе нового сообщения.

Для оптимизации качества кластеризации могут использоваться различные вариации приведенного алгоритма. Так, например, Яндекс.Новости [5], выполняют кластеризацию в несколько проходов с целью объединения атомарных кластеров [4].

За время проведения конференции TDT (с 1998 по 2004 год) было предложено множество достаточно эффективных алгоритмов решения задачи кластеризации новостного потока, однако вопрос о стабильности предложенных методов остается открытым. Как видно из описания алгоритма, эффективность работы системы кластеризации в первую очередь зависит от выбранной меры и правильного определения граничного значения. Несмотря на то, что те или иные граничные значения используется практически во всех работах, нам не известно исследований, направленных на выяснение вопроса стабильности порогового значения на данных коллекции. Для получения повторяемого результата необходимо ответить на следующие вопросы: как ведет себя граничное значение для различных мер? Насколько стабильно найденное граничное значение? Какие меры предпочтительнее в терминах эффективности и стабильности граничных значений?

Целями данного исследования был поиск оптимальных граничных значений для различных мер близости, выяснение стабильности найденного значения и сравнение эффективности алгоритмов.

2 Анализируемые алгоритмы

В современных работах описывается множество различных мер и базирующихся на них алгоритмов кластеризации новостного потока. Для наших исследований были отобраны следующие меры:

2.1 Мера Джаккарда

Наиболее простой мерой близости двух документов представляется мера совпадения множеств термов документов, известная так же как мера Джаккарда (Jaccard) [17]. Эта мера определяется как

$$Sim = \frac{|A \cap B|}{|A \cup B|}$$

где A и B – множества термов, входящих в текстовые документы. Из формулы видно, что мера принимает значения от 0 до 1 и достигает максимума при полном совпадении двух множеств. Данная мера чрезвычайно проста, однако содержит ряд недостатков. Выделим два из них. Во-первых, мера не учитывает разницу в размере сравниваемых документов, а во-вторых, при ее вычислении не используется информация о частоте употребления термов, составляющих документы.

Нами было рассмотрено несколько алгоритмов, использующих эту меру: Story Similarity, Story Minimal Similarity, Subject Similarity, Named Entities Similarity, а также аналогичная мера Sub Similarity.

Алгоритм **Story Similarity** вычисляет расстояние от нового документа до кластера как максимальное значение меры Джаккарда, вычисленное для нового сообщения и документов кластера. **Story Minimal Similarity** так же использует меру Джаккарда, однако в качестве расстояния до кластера принимается минимальное значение меры, вычисленное для нового новостного сообщения и документов кластера. Алгоритм **Subject Similarity** аналогичен Story Similarity, но при вычислении меры использует только информацию заголовков.

В задачах, связанных с обработкой новостного потока и в первую очередь в задачах идентификации новых событий часто применяются методы, анализирующие множества именованных сущностей, присутствующих в документе [6]. Для упрощения задачи в нашем алгоритме **Named Entities Similarity** идентификаторами считались любые термы, начинающиеся с заглавной буквы. В качестве меры близости множеств именованных сущностей использовалась мера Джаккарда.

Для новостных сюжетов характерно, что в ходе развития событий более ранние сообщения агрегируются более поздними. Для проверки этой гипотезы была введена мера **Sub Similarity**, определяющаяся как отношение мощности пересечения множеств термов к мощности наименьшего множества [2]:

$$SubSim = \frac{|(A \cap B)|}{min(|A|, |B|)}$$

Описанная мера показывает, насколько одно новостное сообщение включается в другое. Расстояние от документа до кластера в алгоритме **Substory Similarity** определяется как максимальное значение меры, вычисленное для нового сообщения и документов кластера.

2.2 Голосующая кластеризация

В основе этого алгоритма лежит идея о голосующих классификаторах, успешно показавшая себя в некоторых работах [8], связанных с кластеризацией новостного потока. Мы предполагаем, что использование комбинации классификаторов, каждый из которых вносит вклад в принятие решения, позволит достичь лучших результатов, чем применение классификаторов по одному. В нашем исследовании в качестве голосующих классификаторов использовались вышеперечисленные алгоритмы. Вклад каждого из классификаторов оценивался как

$$VoteVal_i = MeasureVal_i - ThresholdVal_i,$$

где MeasureVal – значение меры схожести i-го алгоритма, ThresholdVal – граничное значение, используемое в алгоритме.

Решение о принадлежности к одному из существующих кластеров принималось на основе анализа суммы значений VoteVal. Новое сообщение относилось к кластеру, для документа которого суммарное значение VoteVal было максимально и превышало заранее установленное граничное значение.

2.3 Tf*Idf мера

Различные вариации мер, основывающихся на векторном представлении документов и tf*idf взвешивании термов чрезвычайно популярны в задачах новостной кластеризации ([11], [15], etc.) В наших экспериментах использовалась широко распространенная мера схожести документов, определяющаяся как косинус угла между векторами, представляющими документы:

$$SimTfIdf = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}$$

где a_i и b_i – компоненты векторов документов A и B, n – размерность векторов. Для построения векторов и вычисления весов термов использовался вариант формулы tf*idf [12]:

$$tf = 0.5 + 0.5 \bullet \frac{TermFrequency}{MaxTermFrequency},$$

где TermFrequency – частота терма в документе, а MaxTermFrequency – максимальная частота термов в документе

$$idf = \log(\frac{N}{df})$$

, где N – общее число документов в коллекции, а df – количество документов, в которых

встречается терм.

Как видно, применение такой меры близости документов позволяет учесть частоты термов документов, однако требует использования статистических данных о частоте термов во всей коллекции. В условиях нашей задачи это требование никогда не может быть выполнено полностью, так как размер коллекции и множество составляющих ее документов заранее неизвестны, однако может аппроксимироваться на основе уже полученных знаний.

2.4 SVM Similarity

Различные исследования, анализирующие методы классификации текстовых документов признают метод опорных векторов (SVM) [9] одним из наиболее эффективных. В нашем исследовании мы применили SVM классификатор для задачи кластеризации новостного потока. В качестве вектора пространства признаков использовался вектор, элементами которого являются значения метрик, полученных при вычислении по алгоритмам Story Similarity, Subject Similarity, Named Entities Similarity, Substory Similarity.

Для обучения классификатора использовалась размеченная новостная коллекция. Обучение выполнялось следующим образом: для каждого нового документа из существующих кластеров последовательно выбирались сообщения, относительно которых строился вектор признаков. В случае если новое новостное сообщение и существующее сообщение должны, согласно разметке, принадлежать различным сюжетам, вектор относился к множеству отрицательных примеров. В случае, когда новое сообщение и существующий документ принадлежат одному кластеру, вектор помечался как положительный пример, из потока выбирался следующий документ и процесс начинался сначала. Выполненное таким образом обучение позволяет свести задачу кластеризации к классификации на два множества. Описанный алгоритм основывается на гипотезе, что вектор признаков содержит достаточно информации для принятия решения о принадлежности (или непринадлежности) истории кластеру.

Алгоритм кластеризации с использованием SVM выглядит следующим образом:

1. Считывается следующее новостное сообщение в потоке
2. Для каждого существующего кластера перебираются входящие в него сообщения, для каждого сообщения кластера строится вектор признаков.
3. Выполняется классификация вектора признаков. В случае, если вектор относится к положительному классу, новое сообщение относится к тому кластеру, к которому принадлежит сообщение, для которого был получен вектор признаков. В противном случае процесс продолжается
4. Если новое сообщение не было отнесено ни к одному из существующих кластеров, на его основе создается новый кластер.

Результат работы приведенного алгоритма сильно зависит от последовательности перебора существующих кластеров и историй в них, так как решение принимается на основе первого положительного примера. В целях устранения этого недостатка нами использовалась модифицированная версия алгоритма, где помимо решения классификатора во внимание принимались значения метрик из пространства признаков. Использование значений метрик позволяет определить кластер, наиболее близкий рассматриваемому сообщению.

3 Эксперименты и настройка граничных значений

3.1 Оценка качества кластеризации

Одной из особенностей задачи кластеризации новостного потока является тот факт, что количество кластеров заранее неизвестно и неограниченно. Эта особенность делает неэффективной широко распространенную оценку качества кластеризации, основанную на понятии энтропии. Из определения метрики следует, что она будет достигать наилучшего значения, когда размер кластеров будет наименьшим [16]. Очевидно, что наилучшее значение при таком методе оценки будет получено при использовании классификатора, создающего новый кластер для каждого нового сообщения потока, что не соответствует действительности. В качестве замены нами использовались следующие две метрики:

Доля верно построенных кластеров (Fr)
Значение данной метрики вычисляется как доля кластеров размеченной коллекции, которая была верно построена системой.

$$Fr = \frac{|A_{correct}|}{|A_{asessed}|}$$

Под верно построенным кластером мы понимаем кластер, состоящий из точно тех же новостных сообщений, что и образец.

F метрика
Видно, что описанная выше метрика не учитывает размер кластеров. Так, например, если в кластере из 100 элементов хотя бы один был приписан к кластеру неправильно, данный кластер не включается в число корректно построенных. Чтобы обойти эту проблему мы ввели F метрику [16], основанную на широко распространенных в задачах информационного поиска метриках точности (p) и

полноты (r). В условиях нашей задачи эти метрики определяются следующим образом:

$P = \frac{|A_{correct}|}{|B|}$, где $A_{correct}$ - множество документов, верно приписанных системой кластеру, а B – кластер-образец.

$R = \frac{|A_{correct}|}{|A|}$, где $A_{correct}$ - множество документов, верно приписанных системой кластеру, а A – все множество документов, приписанных к кластеру системой. Метрика F определяется на основе метрик p и r следующем образом:

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Метрика полагается равной 0, если p или r равны 0. Значения метрики лежат в диапазоне от 0 до 1, причем 1 достигается в случае, когда кластер, построенный системой, полностью совпадает с образцом. Для оценки классификации всей коллекции использовалось среднее значение F.

3.2 Экспериментальная коллекция

Все эксперименты проводились на основе новостной коллекции семинара РОМИП 2005. Для сообщений потока выполнялся стемминг и из текста исключались стоп-слова.

В силу временных ограничений в рамках РОМИП кластеризация данных асессорами не производилась. Для анализа и настройки алгоритмов в нашем исследовании была выполнена ручная разметка двух подмножеств коллекции, состоящих из 500 сообщений каждый (коллекция A и коллекция B). Оценка производилась двумя асессорами, причем первый асессор оценивал коллекцию A, а второй - коллекцию B. В результате выполненной вручную кластеризации в коллекции A были выделены 261, а в коллекции B 268 новостных сюжетов.

Учитывая тот факт, что количество кластеров, выделенных вручную, достаточно велико и среди них значительную часть составляют атомарные кластеры, в качестве базового алгоритма был выбран алгоритм кластеризации, создающий новый кластер для каждого нового сообщения новостного потока. Для базового классификатора были получены следующие значения метрик на первом подмножестве: F=0.83, Fr=0.64.

3.3 Настройка алгоритмов

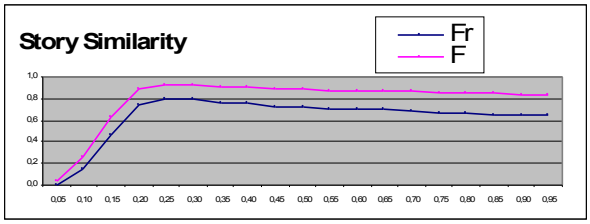
Одной из задач данного исследования было определить оптимальные граничные значения для исследуемых алгоритмов и выяснить устойчивость найденных значений. Исследование алгоритмов проводилось в два этапа. На первом этапе выполнялась настройка граничного значения на первом размеченном подмножестве новостной коллекции.

При настройке алгоритмов выбирались граничные значения в определенном диапазоне с малым шагом. Диапазон выбирался исходя из теоретических ограничений (0..1 для всех алгоритмов кроме голосующей кластеризации). Для каждого из выбранных значений выполнялась кластеризация, результаты которой оценивались с помощью обеих метрик, F и Fr. Рассмотрим результаты экспериментов более подробно.

Алгоритмы, использующие меру Джаккарда

Данная мера использовалась в алгоритмах оценки максимальной и минимальной схожести документов, максимальной схожести заголовков и множеств идентификаторов.

Построение кластера на основе поиска *максимально* похожей истории дает несколько лучшие результаты (максимальные значения метрик F=0.93, Fr=0.8) по сравнению с методом, в котором новое сообщение относится к кластеру, в котором *минимальное значение меры схожести максимально* (максимальные значения метрик F=0.92, Fr=0.77). Полученные результаты подтверждают гипотезу, согласно которой кластер новостных сообщений должен рассматриваться как развивающийся сюжет, и признаком принадлежности кластеру является близость нового документа одному из документов кластера, а не его центру. В то же время, разница в значениях метрик F и Fr составляет 1 и 3 процента соответственно, что не позволяет говорить о принципиальном преимуществе одного из подходов.

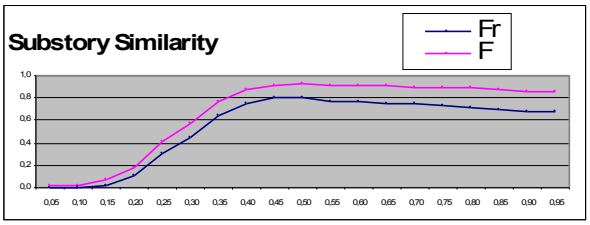


Метод кластеризации, основывающийся на сравнении заголовков, показал достаточно низкие результаты (максимальные значения метрик F=0.88, Fr=0.69). Это позволяет говорить о недостаточности данных, содержащихся в заголовках для принятия решения о принадлежности сюжету.

Метод кластеризации, использующий только идентификаторы, показал наихудшие результаты среди всех рассматривающихся алгоритмов (максимальные значения метрик, F=0.86, Fr=0.69), незначительно превышающие результаты базового классификатора (F=0.83, Fr=0.64). Несмотря на то, что данный алгоритм в некоторой степени не оправдал возложенных на него надежд, может быть

интересно его использование в комбинации с другими методами кластеризации.

Алгоритм, использующий меру SubSim, показал достаточно хорошие результаты, совпадающие с наилучшими результатами описанных выше алгоритмов. Наилучшие значения мер: F=0.93, Fr=0.8

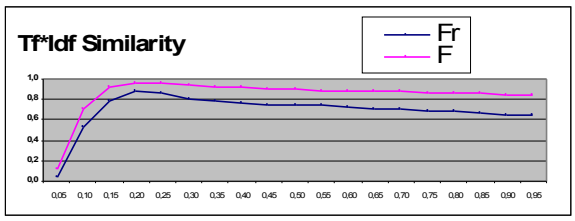


Голосующая кластеризация

Использование комбинации алгоритмов позволило несколько улучшить результат (максимальные значения метрик F=0.93, Fr=0.82) по сравнению со значениями, полученными индивидуальными алгоритмами, участвующими в голосовании.

Если сравнить графики зависимости качества классификации от граничного значения, можно заметить, что все меры, основанные на мере Джаккарда, имеют ярко выраженный промежуток роста, а затем медленно убывают. График голосующего классификатора по форме значительно отличается от зависимостей, полученных для других алгоритмов, в нем нет ярко выраженных промежутков роста и убывания функции. Этот факт объясняется тем, что, комбинируя алгоритмы, мы компенсируем ошибки одних из них правильными решениями других.

Мера Tf * Idf



Использование Tf*Idf меры позволило получить наилучший результат кластеризации (F=0.96, Fr=0.88). Оптимальное пороговое значение составило 0.2. Стоит, однако, заметить, что применение такого метода кластеризации требует накопленной статистической информации о частотах термов в документах новостного потока. В данном эксперименте использовалась статистика, полученная на первом оцененном подмножестве новостной коллекции (словарь составил порядка 11000 слов), к которой затем и применялся алгоритм кластеризации. Такая оценка не совсем корректна, так как в данном случае алгоритм использовал ‘идеальную’ статистику.

Для получения более корректного результата нами был проведен дополнительный эксперимент (см. график), в ходе которого для сбора статистики использовалось порядка 4000 новостных сюжетов (словарь составил 43000 слов). Данная коллекция не включала оцененное подмножество. Результаты эксперимента показали, что оптимальное граничное значение сместилось к отметке 0.25-0.3, значения метрик Fr и F опустились до 0.84 и 0.95 соответственно.

На приведенных графиках можно увидеть, что зависимости качества кластеризации (выраженного в метриках F и Fr) от используемого граничного значения имеют схожую форму и для всех алгоритмов за исключением голосующей кластеризации характеризуются резким ростом при малых граничных значениях с последующим медленным убыванием после достижения максимума. Такой вид графиков позволяет предположить, что для достижения стабильной работы алгоритмов на различных потоках данных необходимо выбирать граничные значения, несколько превосходящие оптимальные. Несмотря на снижение эффективности, такой подход должен минимизировать ошибку.

В следующей таблице приведены оптимальные пороговые значения, найденные для каждого алгоритма и достигаемые на них значения метрик F и Fr.

Алгоритм	Опт. гр. значение	Fr	F
Story Sim	0.25	0.80	0.93
Story Min Sim	0.25	0.77	0.92
Sub sim	0.50	0.80	0.92
Subject Sim	0.25	0.69	0.88
Named entities similarity	0.70	0.69	0.86
Voting Sim	-0.6	0.82	0.93
Tf*Idf Sim	0.30	0.84	0.95

Наши эксперименты показали, что в большинстве случаев обе использованные метрики качества кластеризации имеют одинаковый характер роста, с возможными незначительными отклонениями. В большинстве случаев максимум достигался обеими функциями в одной точке.

3.5 SVM кластеризация

На первом этапе экспериментов основной задачей для SVM кластеризации было обучение классификатора для дальнейшей оценки его эффективности в последующих экспериментах. В результате обучения было получено более 65000 обучающих векторов, большая часть которых относится к множеству отрицательных примеров, то есть описывают случай, когда история не должна быть причислена к рассматриваемому кластеру. В качестве первичной проверки SVM-алгоритма кластеризация была выполнена на обучающем множестве. Полученные результаты превзошли

результаты всех прочих алгоритмов (F=0,95, Fr=0.85) , что позволило признать метод перспективным для дальнейшего анализа.

Подводя итог первого этапа экспериментов по настройке алгоритмов, отметим, что из всех алгоритмов наилучшие значения метрик были достигнуты на алгоритмах Story Similarity, Substory Similarity, голосующей кластеризации и Tf*Idf

4 Второй этап экспериментов

Основными задачами второго этапа экспериментов были проверка стабильности работы исследованных алгоритмов на другом оцененном подмножестве новостной коллекции и сравнение их эффективности. Помимо этого на втором этапе экспериментов проводилась проверка эффективности предложенного нами метода кластеризации на основе SVM классификатора.

Полученные результаты работы базового классификатора для коллекции B: F = 0.83, Fr = 0.63.

Для экспериментов нами были отобраны наиболее перспективные алгоритмы, для которых были поставлены опыты, аналогичные экспериментам первого этапа, но с использованием коллекции B новостных сообщений.

В таблице приведены оптимальные граничные значения, полученные на втором этапе экспериментов и соответствующие им значения метрик качества кластеризации. Для удобства сравнения в таблице так же приводятся граничные значения, оптимальные для первой оцененной коллекции.

Алгоритм	Опт. гр. значение 2	Опт. гр. значение 1	Fr	F
Story Sim	0.25	0.25	0.76	0.91
Sub Sim	0.50	0.50	0.75	0.90
Voting Sim	**-0.40**	**-0.60**	0.75	0.91

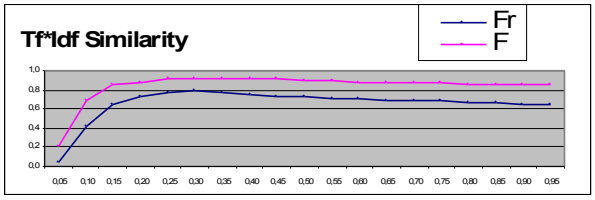
Из приведенных результатов экспериментов видно, что на рассмотренных алгоритмах, за исключением голосующей кластеризации, оптимальное граничное значение достигается в той же точке, что и в экспериментах на первой оцененной коллекции. Дополнительные эксперименты с использованием алгоритмов Subject Similarity и Named Entities Similarity показали, что и в этом случае наблюдается стабильность найденных граничных значений.

Расхождение в оптимальных граничных значениях для алгоритма голосующей кластеризации можно считать условным, так как значения метрик колеблются вокруг среднего значения и максимальная разница по мере Fr, например, достигает всего лишь 5 процентов.

Tf*Idf Similarity

На втором этапе экспериментов Tf*Idf алгоритм применялся 2 раза, в первый раз с использованием статистики, накопленной только на первом оцененном подмножестве коллекции. Результаты алгоритма оказались неожиданно высоки для такого малого объема статистических данных: максимальное значение метрики Fr=0.78, F=0.92. При этом оптимальное граничное значение осталось неизменным по сравнению с опытами первого этапа - 0.3. Видно, что эти результаты превысили результаты работы алгоритмов, не учитывающих частотное распределение термов.

Для сравнения был поставлен второй эксперимент с использованием статистики, полученной на подмножестве коллекции из примерно 4000 документов. Несмотря на больший объем данных, результаты практически не изменились. Лучшее качество кластеризации по-прежнему достигается на отметке 0.3, значения мер Fr и F достигают 0.79 и 0.92 соответственно:



Полученные данные позволяют сделать вывод об относительной стабильности работы рассматриваемых алгоритмов и найденных граничных значений. Важно так же отметить, что вид зависимости эффективности кластеризации от выбранного граничного значения сохраняется на втором подмножестве коллекции.

Несмотря на то, что эксперименты показали относительную стабильность найденных оптимальных параметров алгоритмов, нас интересовал вопрос субъективности оценки коллекции асессором. Для выяснения этого факта Коллекция B была повторно размечена тем же асессором, который оценивал Коллекцию A. Сравнение двух вариантов разметки коллекции B показало, что лишь около 80 процентов кластеров совпадает. Очевидно, что такое сильное расхождение в оценке не могло не сказаться на значениях метрик качества классификации. В связи с полученными данными можно сделать вывод, что результаты работы различных систем существенно субъективны и могут значительно варьироваться в зависимости от конкретного асессора.

4.1 SVM Similarity

Одной из задач второго этапа экспериментов была проверка эффективности применения SVM классификатора для решения задачи кластеризации

новостного потока. SVM классификатор, обученный на коллекции A, показал средние результаты при кластеризации коллекции B: Fr=0.76, F=0.90.

Тот факт, что обучение даже на небольшой коллекции (500 документов) позволило достичь достаточно хорошего результата на независимом множестве документов, позволяет расценивать SVM метод как перспективный алгоритм, требующий более подробного изучения. В дальнейших исследованиях мы планируем изучить влияние расширения пространства признаков и конкретного ядра алгоритма на эффективность кластеризации.

5 Заключение

В данной работе был рассмотрен ряд алгоритмов кластеризации новостного потока и определены оптимальные параметры данных алгоритмов с использованием размеченного подмножества новостной коллекции РОМИП 2005. Наилучшие результаты по итогам двух серий экспериментов были показаны алгоритмом Tf*Idf.

Проведенные исследования поведения различных алгоритмов подтвердили относительную стабильность оптимального граничного значения при инкрементальной кластеризации. Предложенный нами метод кластеризации на основе классификатора SVM продемонстрировал достаточно хорошие результаты, однако требует дополнительных исследований и оптимизации.

Литература

1. Добров Б.В., Лукашевич Н.В., Штернов С.В., Обработка потока новостей на основе больших лингвистических ресурсов. Сборник работ научных стипендиатов Яндекс Интернет-Математика 2005, Ярославль, 2005.
2. Зевайкин А.Н., Корнеев В.В., Формирование выпуска новостей на основе автоматического анализа новостных сообщений. Сборник работ научных стипендиатов Яндекс Интернет-Математика 2005, Ярославль, 2005.
3. Российский семинар по Оценке Методов Информационного Поиска. http://romip.narod.ru/
4. Сегалович И., Маслов М., Нагорнов Д., Как работают новые Яндекс.Новости. http://company.yandex.ru/technology/publications/2003-08.xml
5. Яндекс.Новости. http://news.yandex.ru/about.html
6. Abdul-Jaleel N., Allan J., Croft W., Diaz F, Larkey L., Li X., Metzler D., Smucker D., Strohman T., Turtle H., Wade C., UMass at TREC 2004: Notebook. In E. Voorhees, editor, The Thirteenth Text Retrieval Conference (TREC 2004) Notebook, pages 657--670, 2004.
7. Allan J., Introduction to topic detection and tracking. James Allan, editor, Topic detection and Tracking: Event-based Information Organization, pages 1-16. Kluwer Academic Publishers, Boston, 2002.
8. Braun R. K., Kaneshiro R., Exploiting Topic Pragmatics For New Event Detection In TDT-2004. DARPA Topic Detection and Tracking Workshop, Gaithersburg, December 2004.
9. Burges C.J.C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):955-974, 1998.
10. Cohen W., Ravikumar P., Fienberg S., A comparison of string distance metrics for name-matching tasks. In Proceedings of the IIWeb Workshop at the IJCAI 2003 conference, 2003.
11. Connel M., Feng A., Kumaran G., Raghavan H., Shah C., Allan J., Umass at TDT2004. Proc. DARPA Topic Detection and Tracking Workshop, Gaithersburg, December 2004.
12. Greengrass E., Information retrieval: A survey. DOD Technical Report TR-R52-008-001, 2001
13. Google News. http://news.google.com/
14. Lo Y.Y., Gauvain J.L., The LIMSI Topic Tracking System for TDT2001. Proc. DARPA Topic Detection and Tracking Workshop, Gaithersburg, November 2001.
15. Schultz J.M. Liberman M., Topic Detection and Tracking using idf Weighted Cosine Coefficient. Proceedings of the DARPA Broadcast News Workshop, 189-192, 1999.
16. Steinbach M., Karypis G., and Kumar V., A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.
17. Strehl A., Ghosh J., Mooney R., Impact of similarity measures on web-page clustering. In Proc. AAAI Workshop on AI for Web Search (2000), 58-64, 2000.
18. Walls F., Jin H., Sista s., Schwartz R., Topic Detection in Broadcast news. Proceedings of the DARPA Broadcast News Workshop, 193-198, 1999.

Analysis of some methods of the automated news clusterization

In this paper we analyze some of the news clusterization algorithms and provide description of the experiments performed to figure out the dependencies between thresholds used, clusterization quality and stability