Исследование качества базовых методов кластеризации новостного потока в суточном временном окне

© Б.В. Добров[1,2], А.М. Павлов[2,3]

[1]Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова, г. Москва

[2]АНО Центр информационных исследований, г. Москва

[3]Институт спектроскопии РАН, г. Москва

dobroff@mail.cir.ru, apavlov_86@mail.ru

Аннотация

Рассматривается задача оценки методов кластеризации новостного потока текстовых документов. Исследуется несколько базовых методов кластеризации, качество которых оценивается по разным метрикам относительно «золотого стандарта» (вручную выполненной разметки на кластеры) распределения новостных кластеров по трем дням новостной коллекции РОМИП 2006 (фрагмент архива ресурса Яндекс.Новости 2003 – 2004 гг.). Проведенные эксперименты показали, что рассмотренные базовые методы показывают близкие результаты.

1 Введение

Новостная информация – сообщения информационных агентств, документы средств массовой информации – являются одними из самых востребованных среди пользователей интернета, аналитических служб корпораций.

Новостные агрегаторы, которые интегрируют новости из тысяч источников в «новостные кластеры», предоставляя пользователям доступ к ранжированным по значению и тематики сюжетам, являются одними из самых популярных ресурсов.

В Рунете широко известны такие новостные агрегаторы, как (в алфавитном порядке): Google.Новости (news.google.ru), Новотека (novoteka.ru), Рамблер.Новости (news.rambler.ru), Яндекс.Новости (news.yandex.ru). В данных информационных ресурсах реализуются разные стратегии представления новостной информации пользователям.

Ключевой проблемой новостных агрегаторов является решение задачи кластеризации – формирование групп близких документов, моделирующих новостные сюжеты. На практике результат автоматической кластеризации может быть затем изменен в зависимости от корпоративной политики (в том числе и вручную редакторами), данных о запросах пользователей или обсуждений в блогах.

Используемые на практике методы кластеризации зависят от большого числа параметров, настроенных на реальные новостные потоки.

Несмотря на огромное количество предложенных методов кластеризации [4, 5, 7, 18, 22, 3], в том числе для кластеризации документов, отсутствуют публикации по тестированию качества предложенных методов на доступных коллекциях новостных сообщений.

Известны результаты оценки методов кластеризации на коллекциях новостных сообщений Reuters [20, 21], где рассматривается тематическая кластеризация, в то время как в новостных агрегаторах производится событийная кластеризация, когда документы группируются вокруг некоторого события.

В настоящей работе мы приводим полученные нами результаты для нескольких базовых методов кластеризации на данных, доступных в рамках участия в программе РОМИП (Российского семинара по Оценке Методов Информационного Поиска) фрагментов коллекций Яндекс.Новости за 2003 – 2004 гг. (так называемая новостная коллекция РОМИП 2006 года).

2 Новостные агрегаторы

2.1 Задачи новостных агрегаторов

Новостные агрегаторы представляют собой сложные программно-аппаратные комплексы, решающие большой круг разнообразных задач. К основным задачам новостных агрегаторов относятся:

- собственно, кластеризация;
- ранжирование документов внутри кластера, включая определение первоисточников информации, перепечаток, определения новизны, актуальности и т. п.;
- обзорное реферирование;
- выявление основных действующих лиц, прямой и косвенной речи;
- ранжирование кластеров по их важности;

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

- формирование и обновление главной страницы сервиса (значительно более посещаемой, чем остальные), определение «главных сюжетов»;
- тематическая классификация, для формирования тематических разделов, рассылок;
- поиск по новостям, обычно, результатом является поиск по кластерам, если хотя бы один документ кластера релевантен запросу;
- обмен данными с другими сервисами портала, например, участие в определении типа поискового запроса как новостного;
- и т. д.

Следует иметь в виду, что все время поступают новые новостные документы, а вчерашние [16, 8] «устаревают».

Также существуют достаточно жесткие требования на время обработки новой порции документов и внесения изменений в текущее распределение новостей.

Задачи кластеризации решаются во временном окне величиной от одних до нескольких суток. При этом для (достаточно редких) длительных новостных сюжетов, более длительных, чем размер временного окна, обычно производится «присоединение» текущего кластера к хранимому в базе данных.

В данной работе мы будем рассматривать суточное временное окно в течение 24 часов одного дня.

2.2 Особенности новостного потока

Существуют определенные представления, как должны выглядеть «идеальные» новости, что отражается во многих книгах [2, 6, 8] в инструкции, как «надо писать новости».

Считается [6], что для содержимого новости должна быть справедлива формула, приписываемая еще римскому ритору Квинтилиану – (кто сделал? + что сделал? + какими средствами? + зачем? + когда? + где?). В англоязычной интерпретации – закон «пять W и одно H» – (Who? + What? + Where? + Why? + When? + How?), приписываемый Р. Киплингу [8].

«Идеальное» новостное сообщение желательно должно быть посвящено одному событию. При этом структура «идеального» новостного сообщения должна иметь следующий вид:
- заголовок, который должен быть максимально информативным;
- основное содержание – не более одного-двух абзацев;
- детализация и объяснение главной мысли сообщения;
- историческая справка;
- энциклопедическая справка.

В табл. 1 приведен пример такого правильно выстроенного новостного сообщения.

Если бы все новостные сообщения строились по единой структуре, то решение задачи кластеризации могло бы значительно упроститься. К сожалению, далеко не все сообщения соответствуют идеальным представлениям.

Таблица 1. Пример правильно выстроенного новостного сообщения

2009-10-05 19:41:34 AK&M	
заголовок	А.Чубайс вошел в список лиц, причастных к аварии на СШГЭС
основное содержание	Экс-глава РАО «ЕЭС России» Анатолий Чубайс назван одним из шести человек, которые, по мнению экспертов Ростехнадзора, были причастны к созданию условий аварии на Саяно-Шушенской ГЭС.
подробности	Об этом говорится в … Кроме того, … Также …
история	Напомним, авария на Саяно-Шушенской ГЭС произошла 17 августа.
справка	Саяно-Шушенский гидроэнергетический комплекс расположен на реке Енисей на юго-востоке Республики Хакасия в Саянском каньоне – у выхода реки в Минусинскую котловину. Комплекс включает Саяно-Шушенскую ГЭС и расположенный ниже по течению контррегулирующий Майнский гидроузел.

Кроме того, «идеальное» новостное сообщение неявно соответствует идеальному сюжету, который отслеживает эволюционно развивающееся идеальное событие, или агрегирует разные подробности одного события (см. табл. 2).

Таблица 2. Пример «идеального» сюжета для кластеризации

μ	Время публикации	Заголовок	Источник
0,39	2009.10.05 16:08:54	Игрок сборной Аргентины забил головой с сорока метров	YTPO.ru
0,41	2009.10.05 16:16:00	Футболист забил головой с сорока метров	ИА «Курсор»
0,40	2009.10.05 16:26:00	Аргентинский форвард забил победный гол ударом головой с середины поля (ВИДЕО)	NEWSru.com
0,48	2009.10.05 16:51:11	Аргентинский футболист забил гол ударом головой с 40 метров	Energyland
0,43	2009.10.05 18:29:00	Удар головой с 40 метров завершился голом (видео)	Футбол. Плюс. Хоккей
1,00	2009.10.05 19:57:16	Аргентинский футболист забил гол ударом головой с 40 метров (видео)	Футбол России

Столбец «μ» отражает оценку близости документов к «главному» документу сюжета, точнее функция $\mu(.)$ определена в разделе 3.

В реальности обычно часто в рамках новостного сюжета одновременно происходит несколько связанных событий, причем один из источников счита-

ет более значимым одно из них, кратко упоминая о других, другие источники вольны поступать наоборот. В результате, образуется сложная структура сюжета, в котором сложно взаимодействуют разные события и их «идеальное» разделение становится нетривиальной задачей (см. табл. 3).

Таблица 3. Пример сложного новостного кластера, «неоптимально» построенного одним из методов, описываемых в разделе 4

μ	Время публикации	Заголовок	Источник
0,24	09.10.05 16:16:00	Шпилька в бок / Износ 98%? – продолжаем работать! :: Общество	chaskor.ru
0,30	09.10.05 19:15:27	Виновников трагедии на ГЭС назовет СКП	YTPO.ru
0,40	09.10.05 19:33:32	Ростехнадзор утвердил методику проверок ГЭС	Ведомости – лента новостей
0,27	09.10.05 19:37:00	Дело об аварии на ГЭС передано Главному управлению СКП РФ	ПРАВО.RU
0,37	09.10.05 19:42:56	Списки лиц, ответственных за аварию на СШГЭС, могут быть расширены.	РБК. Главные новости
0,28	09.10.05 19:52:32	Ростехнадзор утвердил методику проверок ГЭС	Голос России – новости
0,20	09.10.05 19:54:11	Секреты долгожителя Чубайса	Svobodanews.ru
0,45	09.10.05 20:06:14	Ростехнадзор продолжит изучать последствия аварии на СШГЭС	Деловая газета «Взгляд»
0,61	09.10.05 20:36:04	Ответственных за аварию на Саяно-Шушенской ГЭС станет больше	MIGnews.com.ua – Украина
1,00	09.10.05 20:51:33	Ростехнадзор пообещал расширить список виновных в аварии на Саяно-Шушенской ГЭС	РегКорреспондент.net – Украина – Россия

3 Формализация задачи кластеризация документов

Рассмотрим задачу кластеризации документов более формально.

Пусть имеются множество документов

$$D = \{d_1, d_2, \ldots, d_N\}$$

и распределение кластеров

$$\Pi = \{C_1, C_2, \ldots, C_K\},$$

такие, что любой из документов $d_i \in C_k$, i=1, .., N, k=1, .., K, и

$$D = \bigcup_k C_k .$$

Мы будем рассматриваем так называемую «жесткую» кластеризацию:

$$C_k \cap C_l = \varnothing \text{ для любых } k, l = 1, \ldots, K.$$

В данной работе для решения задачи кластеризации новостного потока мы будем рассматривать метрические алгоритмы кластеризации [7], учитывающие особенности новостных полнотекстовых документов, т.е. произведем векторизацию задачи – представим документы в виде векторов в метрическом пространстве и введем меру близости.

3.1. Мера близости между документами

Введем меру близости между документами, в качестве которой будем рассматривать косинус между нормализованными векторами признаков документов:

$$\mu\ (di, dj) = \sum_m \frac{d_{im}}{\|d_i\|} \cdot \frac{d_{jm}}{\|d_j\|} .$$

Матрица M=$\{\mu_{ij}\}$ = $\{\mu\,(d_i,d_j)\}$ называется матрицей близости между документами.

Собственно содержимое таблиц 2 и 3 наглядно иллюстрирует основную проблему метрических методов кластеризации. При значении косинусовой меры близости более некоторого порога (например, 0,40) – документы заведомо близки. При значении меры близости меньше другого порога (примерно 0,28) реальная семантическая близость между документами (оцениваемая экспертами) может вступать в противоречие с формальной мерой близости.

3.2 Мера близости между кластерами

Меру близости между разными кластерами, как частный случай, близость между отдельным документом и кластером, можно определить разными способами.

1. Расстояние ближнего соседа:

$$\mu_{\max}(C_k, C_l) = \max_{\substack{d_i \in C_k \\ d_j \in C_l}} \mu\,(d_i, d_j) .$$

2. Расстояние дальнего соседа:

$$\mu_{\min}(C_k, C_l) = \min_{\substack{d_i \in C_k \\ d_j \in C_l}} \mu\,(d_i, d_j) .$$

3. Групповое среднее расстояние:

$$\mu_{average}(C_k, C_l) = \frac{1}{|C_k|} \cdot \frac{1}{|C_l|} \cdot \sum_{\substack{d_i \in C_k \\ d_j \in C_l}} \mu\,(d_i, d_j) .$$

4. Групповое расстояние между нормализованными центрами кластеров:

$$\mu_{center}(C_k, C_l) = \mu(d_*(C_k), d_*(C_l)),$$

где, например,

$$d_*(C_k) = \sum_{d_i \in C_k} \frac{d_i}{\| \sum_{d_j \in C_k} d_j \|},$$

$d_*(Cl)$ определяется аналогично.

3.3 Векторизация документов

3.3.1 Морфологические индексы

Документы подвергались морфологическому анализу.

Вес леммы для данного документа вычисляется по формуле $TF \cdot IDF$ [15, 1], которая учитывает частоту вхождения слова в документ и количество документов коллекции, содержащих данное слово. Мы используем вариант широко распространенный формулы $TF \cdot IDF - BM25 \cdot INQUERY$ [1] – для леммы (нормализованной словоформы) λ документа d:

$$TF \cdot IDF_d(\lambda) = \beta + (1-\beta) tf_d(\lambda) idf(\lambda),$$

$$tf_d(\lambda) = \frac{freq_d(\lambda)}{freq_d(\lambda) + 0{,}5 + 1{,}5 \cdot \frac{dl(d)}{avg\_dl}},$$

$$idf(\lambda) = \frac{\log\left(\frac{|c| + 0{,}5}{df(\lambda)}\right)}{\log(|c| + 1)},$$

где $tf_d(\lambda)$ (*term frequency*) – учет частотности леммы в документе; $freq_d(\lambda)$ – частотность леммы l в документе, dl_d – мера длины документа (количество разных лемм), avg_dl – средняя длина документа, β=0.4;

$idf(\lambda)$ (*inverse term frequency*) – фактически форма штрафования часто используемых в коллекции слов, $|c|$ – количество документов в коллекции, $df(\lambda)$ – количество документов, где встретилось лемма λ.

Слова, встретившиеся в документе, получают ненулевой вес, вес остальных слов равен нулю.

Морфологический индекс формуле TF*IDF обозначим через *L*-индекс.

Аналогичным образом формировался индекс по заголовкам документов, который мы обозначим через *H*-индекс.

3.3.2 Концептуальный индекс

Также документы обрабатывались программным обеспечением АЛОТ (Автоматизированная Лингвистическая Обработка Текстов) [9], когда для документа производится:

- терминологический анализ (выявление терминов Общественно-политического тезауруса [10]), в том числе разрешение многозначности;
- тематический анализ, формирование тематического представления [9], то есть определение основных и второстепенных тематически связанных групп понятий тезауруса, что позволяет определить для каждого понятия текста его вес в зависимости от места в тематическом представлении.

Индекс по понятиям тезауруса мы обозначим как *C*-индекс.

3.3.3 Модификация индексов

Использовались следующие модификации индексов (настроенные ранее на реально работающем приложении кластеризации новостного потока):

- использовались словари «стоп-слов», в том числе для понятий тезауруса, которые исключались из индексов;
- слова с большой буквы получали несколько больший вес (вес в соответствующем индексе умножался на множитель 1,3);
- использовался специальный индекс, немного повышающий значимость географических названий (множитель 1,1), а также понижающий вес «кросс-кластерных» слов и понятий (множитель 0,3) – например, для слов «ДТП», «биржа», «ветер» и т. д.

Для каждого из *L*-, *H*-, *C*-индексов бралось не более 20 элементов с максимальным весом, при этом в качестве окончательной оценки близости рассматривалась мера:

$$\mu_{LCH}(d_i, d_j) = \alpha_L \mu_L(d_i, d_j) + \alpha_C \mu_C(d_i, d_j) + \alpha_H \mu_H(d_i, d_j),$$

где $\alpha_L + \alpha_C + \alpha_H = 1$. В дальнейшем, для обозначения конкретного выбора способа *LCH*-векторизации используется обозначение вида 50:20:30, что соответствует α_L=0,50, α_C=0,20, α_H=0,30.

4 Базовые методы кластеризация

В данной работе мы рассмотрим следующие известные методы кластеризации:

- метод агломеративной кластеризации;
- метод *k*-средних;
- метод FOREL;
- метод DBSCAN.

Опишем данные методы.

4.1 Метод агломеративной кластеризации

В методе агломеративной кластеризации [15] постепенно объединяются наиболее близкие кластеры, начиная с отдельных документов.

Краткое описание:

- сначала каждый документ образует свой кластер;
- до тех пор, пока существуют два достаточно близких кластера

$$\mu(C_i, C_j) \geq \text{Config.Threshold},$$

где Config.Threshold – параметр метода, самые близкие друг к другу кластеры объединяются.

Для расчёта близости двух кластеров используем следующий алгоритм:

- для одиночных кластеров (кластеров, состоящих из одного документа) в качестве меры близости между ними берём близость их единственных документов;
- для других случаев рассчитываем меру по формуле Ланса – Вильямса [7, 17], позволяющей при соответствующих значениях параметров выбирать один из видов мер близости (см. п. 3.1.2).

Результатом работы метода агломеративной кластеризации становится иерархия объединяющихся кластеров. Задание параметра Config.Threshold «разрезает» иерархию на итоговое распределение кластеров.

4.2 Метод *k* средних

Алгоритм метода *k* средних (*k means*) [15] заключается в том, что:

1) фиксируется *k* центров кластеров;
2) все документы кластеризуются – относятся к ближайшему центру;
3) затем производится пересчет центров кластеров;
4) метод останавливается либо по количеству итераций, либо по сходимости изменения движения центров кластеров.

Классический метод *k* средних допускает центры кластеров, не совпадающие с каким-либо из документов. Мы рассматривали модификацию метода *k* средних, так называемый метод *k* центров (*k medoids*), когда на очередной итерации новым центром кластера становится один из документов коллекции.

4.3 Модифицированный метод *k* средних

Метод *k* средних очень прост, допускает введение различных модификаций, но в классической версии требует изначального задания числа кластеров *k* и может приводить к зацикливанию, а также плохо обрабатывает удаленные точки («outliers»). Поэтому вместо классического метода *k* средних мы рассматривали его модифицированный аналог:

- количество кластеров не фиксируется, а определяется путем грубой оценкой на первой итерации:
  - если для непросмотренного документа найдутся кластеры, центры которых находятся ближе, чем порог Config.ClusterFirstThreshold, то новый документ присоединяется к ближайшему кластеру;
  - иначе – он образует новый кластер и становится его центром;
- далее производится конечное количество итераций:
  - делается шаг по алгоритму *k* среднего, но документ относится к кластеру только в случае, если мера близости больше порога Config.ClusterFirstThreshold;
  - производится дополнительная очистка кластера (операция Remaining) – после пересчета центра кластера из него удаляются «далекие» документы, которые имеют меру близости с центром менее порога Config.ClusterRemainingThreshold;
- дополнительно производится объединение кластеров (операция Glue), центры которых имеют между собой меру близости, большую порога Config.ClusterGlueThreshold, после этого применяется завершающая операция Remaining с порогом Config.ClusterRemainingThreshold2.

Основная идея разработки модифицированного метода *k* средних – ввести дополнительные параметры «свободы», настройка которых позволила бы точнее отслеживать специфику задачи кластеризации новостного потока (отслеживание актуальности, перепечаток и т. п.).

4.4 Метод FOREL

Метод FOREL (ФОРмального ЭЛемента) предложен Н.Г. Загоруйко и В.Н. Ёлкиной [11, 7].

Краткое описание алгоритма:

- все документы помечаются как непросмотренные;
- до тех пор, пока есть непросмотренные документы:
  - берём первый непросмотренный документ и делаем его новым кластером;
  - повторяем следующие итерации, пока кластер не перестанет изменяться;
    - строим центр кластера через усреднение векторов документов;
    - находим все близкие (с мерой близости больше заданного порога Config.ClusterThreshold) к центру кластера документы, помеченные как непросмотренные, и делаем их документами нашего кластера;
    - фиксируем новый кластер, помечаем его документы как непросмотренные.

4.5 Метод DBSCAN

Метод DBSCAN (Density-Based Spatial Clustering of Applications with Noise) относится к методам кластеризации по плотности элементом множества (density clusterization) [17].

Неформальное описание:

- все документы помечаются как непросмотренные;

- до тех пор, пока есть непросмотренные документы:
  - берём первый непросмотренный документ и находим ближайшие к нему;
  - если есть достаточное количество близких документов, то рассматриваемый документ образует новый кластер, в который также входят близкие документы (при условии, что они не вошли до сих пор в другой кластер), и к новым документам рекурсивно применяется та же процедура;
  - если близких документов недостаточно, документ считается шумом и помечается, как просмотренный.

  Подробное описание.
- Пусть: *docset* – множество всех документов, *noiseset* – пустой список;
- В цикле: пока *docset* не пусто:
  - берём первый документ *doc* из *docse*t (таким образом удаляем его из *docset*);
  - ищем документы с близостью к *doc* не меньше, чем Config.ClusterThreshold;
  - если их количество меньше, чем Config.MinNumber, то помещаем *doc* в *noiseset*;
  - иначе создаём новый кластер $C_i = \{doc\}$:
    - создаём *workset* и помещаем в него близкие к *doc* документы, содержащиеся в *docset* (с удалением из *docset*);
    - близкие к *doc* документы, содержащиеся в *noiseset,* помещаем в C_i (с удалением из *noiseset*);
    - пока *workset* не пусто:
      - берём первый *docj* из *workset*, удаляем его из *workset*, добавляем в C_i и ищем близкие к нему документы;
      - если их количество не меньше, чем Config.MinNumber, переносим те из них, что содержатся в *noiseset* в C_i, а те, что содержатся в *docset,* – в *worklist*.

Метод DBSCAN имеет простую интерпретацию. Если представить множество документов в виде графа, когда ребро между вершинами графа (документами) проводится при условии, что мера близости между ними не меньше, чем Config.ClusterThreshold. Тогда в кластер отбираются те вершины графа, которые имеют не менее заданного количества ребер с уже отобранными вершинами.

В своей работе мы немного модифицировали алгоритм – после его окончания мы итерационно применяли тот же метод к оставшимся документам, уменьшая каждый раз порог по количеству ребер.

5 Эксперимент по оценке методов кластеризации

Оценка методов кластеризации традиционно считается сложной проблемой [4, 12 – 14]. Наиболее доверенные оценки качества могут быть получены сравнением результатов кластеризации с коллекциями, размеченными вручную. Проведение ручной разметки кластеров считается очень трудоемкой задачей.

Известны работы, когда оценка кластеризации выполняется сравнением с результатами классификации, когда каждому документу приписывается ровно одна рубрика, что, например, практически выполняется для коллекции Reuters-21578. К сожалению, для задачи кластеризации полного новостного потока таких коллекций не известно. Тем не менее, авторы поставили и решили задачу создания «золотого стандарта» для рассматриваемой задачи, что позволило получить реальные оценки качества рассматриваемых методов кластеризации.

5.1 Тестовая коллекция

Мы рассматривали новостную коллекцию версии 2006 года, доступную по программе РОМИП (Российского семинара по Оценке Методов Информационного Поиска,), содержащую подколлекции новостного агрегатора Яндекс.Новости за три недели 2003 – 2004 гг. (http://romip.ru/ru/collections/news-collection.html).

Для оценки качества кластеризации новостей в суточном окне мы рассмотрели три дня, являющимися средами соответствующих недель (см. табл. 4).

Таблица 4. Характеристики дней новостной коллекции, отобранных для оценки

Недели	Дни	Количество документов
Неделя Шеварднадзе	2003-11-20	1752
Обычная неделя	2003-12-03	1715
Неделя выборов	2004-04-02	1809

5.2 «Золотой стандарт»

Для каждого из рассматриваемых дней одним из методов кластеризации было получено «приближенное» распределение кластеров. Затем с помощью специального программного обеспечения редактирования распределения кластеров авторы независимо формировали «идеальное» распределение.

Программное обеспечение редактирования позволяет эффективно выполнять следующие операции:

- визуализировать кластеры – как упорядоченный список, сортированный по размеру кластеров; для каждого кластера просмотреть его состав в документах, сортируя из по времени публикации, близости к центру; для каждого документа можно просмотреть текст;
- для каждого кластера выводятся также близкие кластеры (кандидаты на «склейку»);
- можно объединять («склеивать»), разделять существующие кластеры.

На практике оказалось, что после некоторого периода освоения технологии редактирования распределения кластеров трудоемкость разметки одного дня для новостной коллекции РОМИП составляло 3 – 4 человека-часа.

5.3 Метрики сравнения

Теоретическое сравнение различных метрик оценки методов кластеризации приведено в работе [19]. Мы для сравнения двух распределений использовали *F1*-меру по парам документов.

Пусть:

- N11 – количество пар документов, таких, что и в эталонном, и в исследуемом распределениях пара в одном кластере;N00 – количество таких пар, что и в эталонном, и в исследуемом распределении пара – в разных кластерах;
- N10 – количество таких пар, что в эталонном распределении пара документов – в одном кластере, а в исследуемом распределении – пара в разных кластерах;
- N01 – количество таких пар, что в эталонном распределении пара документов – в разных кластерах, а в исследуемом распределении пара – в одном кластере.

Тогда:

точность кластеризации

$$P = Precision = N11 / (N11 + N01),$$

полнота кластеризации

$$R = Recall = \mathrm{N11} / (\mathrm{N11} + \mathrm{N10}),$$

F1-мера

$$F1 = 2 \cdot P \cdot R / (P + R).$$

Планировалось проводить усреднение по «идеальным» распределениям, полученным при сравнении методов кластеризации разными экспертами. Однако оказалось, что при сравнении «идеальных» распределений разных экспертов *F1*-мера превышала 95 %, поэтому результаты сравнивались только с одним из «идеальных» распределений.

5.4 Сравнение разных методов

Для каждого из методов мы подбирали лучшие параметры *LCH*-векторизации (см. п. 3.3.3), метод вычисления меры близости между кластерами (см. п. 3.2.1), пороги Config.*Threshold методов. Среднее время полного расчета по одному методу с выбранным набором параметров составляло около трех минут, включая построение индексов.

Результаты по сравнению методов, описанных в разделе 4, приведены в табл. 5. Лучшие результаты по каждому дню и усреднено по трем дням выделены полужирным шрифтом.

Оказалось, что результаты полученные для разных методов, достаточно близки.

Таблица 5. Сравнение различных базовых методов кластеризации по трем дням отдельно и усредненному (Result – наилучшее значение метрики *F1*, Ratio – отношение лучшего результата к рассматриваемому, *LCH* – способ векторизации, Method – метод расчета связей, Threshold – параметры радиус близости)

Метод	2003-11-21	2003-12-03	2004-04-02	Среднее
FOREL/ *LCH* = 60:20:20	Result = 0,5282 Ratio = 1,092 Method = center Threshold = 0,32	Result = 0,8383 Ratio = 1,036 Method = center Threshold = 0,34	Result = 0,7364 Ratio = 1,073 Method = average Threshold = 0,24	Result = 0,6890 Ratio = 1,051 Method = center Threshold = 0,34
DBSCAN/ *LCH* = 60:20:20	Result = 0,5173 Ratio = 1,115 Number = 8 Threshold = 0,28	Result = 0,8648 Ratio = 1,004 Number = 5 Threshold = 0,30	Result = 0,7504 Ratio = 1,053 Number = 3 Threshold = 0,32	Result = 0,6879 Ratio = 1,053 Number = 5 Threshold = 0,30
Modified K-Means/ *LCH* = 60:15:25	**Result = 0,5767** **Ratio = 1,000** **Iterations = 0,26** **Remaining = 0,20** **Remaining2 = 0,06** **Glue = 0,30**	Result = 0,8515 Ratio = 1,020 Iterations = 0,22 Remaining = 0,22 Remaining2 = 0,10 Glue = 0,28	Result = 0,7616 Ratio = 1,038 Iterations = 0,22 Remaining = 0,22 Remaining2 = 0,06 Glue = 0,32	Result = 0,7141 Ratio = 1,014 Iterations = 0,24 Remaining = 0,22 Remaining2 = 0,06 Glue = 0,32
Agglomerative/ *LCH* = 60:15:25	Result = 0,5470 Ratio = 1,054 Method = center Threshold = 0,26	Result = 0,8250 Ratio = 1,053 Method = average Threshold = 0,18	Result = 0,7549 Ratio = 1,047 Method = min Threshold = 0,30	Result = 0,7003 Ratio = 1,034 Method = center Threshold = 0,26
Agglomerative	Result = 0,5716 Ratio = 1,008 *LCH* = 40:40:20 Method = average Threshold = 0,22	**Result = 0,8685** **Ratio = 1,000** ***LCH* = 40:30:30** **Method = center** **Threshold = 0,32**	**Result = 0,7904** **Ratio = 1,000** ***LCH* = 20:50:30** **Method = center** **Threshold = 0,34**	**Result = 0,7243** **Ratio = 1,000** ***LCH* = 40:30:30** **Method = center** **Threshold = 0,30**

Таблица 6. Сравнение эффективности использования различных способов векторизации для метода кластеризации Agglomerative

LCH = (0, 0, 100)	*LCH* = (100 ,0, 0)	*LCH* = (*x*,0,100-*x*)	*LCH* = (*x*,*y*,100-*x*-*y*)
Result =0,4972 Ratio = 1,457	Result =0,5767 Ratio = 1,256	Result =0,6866 Ratio = 1,055	**Result = 0,7243** **Ratio = 1,000**
		LCH = 70:00:30	***LCH* = 40:30:30**
Method = center Threshold = 0,38	Method = min Threshold = 0,38	Method = center Threshold = 0,26	**Method = center** **Threshold = 0,30**

Несколько лучшие результаты показали метод агломеративной кластеризации и модифицированный метод *k* средних.

Отметим, что наблюдается достаточно существенный разброс в качестве кластеризации между разными днями – от 57,7 % *F1*-меры для дня 2003-11-21 до 86,9 % для дня 2003-12-03.

Среднее по трем дням составляет 72,4 % по *F1*-мере.

5.5 Влияние выбора способа векторизации

В табл. 6 приведены результаты исследования влияния выбора *LCH*-векторизации на качество кластеризации для метода агломеративной кластеризации (при усреднении по трем дням).

Если пытаться кластеризовать документы только по заголовкам, то будут получены достаточно низкие результаты (F1=49,7%), что в определенной мере подтверждает обоснованность выбранной метрики.

Используя только результаты морфологического анализа, особое выделение заголовков (лучшая векторизация *LCH* = 70:00:30) дает результат *F1*=68,7% против *F1*=57,7% для случая без выделения заголовков (улучшение 19%).

Применение тезауруса и тематического представления улучшает результаты еще на 5,5% до *F1*=72,4%.

6 Выводы

Мы реализовали несколько базовых методов кластеризации, для которых выполнили подбор оптимальных параметров для новостных подколлекций Яндекс.Новости 2003 – 2004 гг. по трем дням (в среднем 1750 документов в день).

Проведенные эксперименты по оценке методов кластеризации на трех «суточных» подколлекциях новостной коллекции РОМИП 2006 показали:

- трудоемкость создания «золотого стандарта» при наличии несложного специального программного обеспечения не является чрезмерной; отметим, что это обстоятельство может влиять на один из основополагающих принципов семинара РОМИП при сравнении методов разных систем – невозможность ручной «подкрутки» результатов – для доверительной оценки качества кластеризации новостного потока надо использовать коллекции большего размера;
- «идеальные» коллекции, составленные независимо разными экспертами, очень близки; таким образом, показана возможность проведения прямой оценки качества методов кластеризации сравнением с «золотым стандартом»;
- все основные методы кластеризации показывают примерно одинаковые результаты при соответствующих оптимальных наборах параметров; в наших экспериментах метод агломеративной кластеризации и модифицированный метод *k* средних показали немного лучшие результаты;
- в среднем качество по *F1*-мере по парам документов составляет 72,4%, при этом для одного из дней 57,7%.

К сожалению, исследуемые коллекции РОМИП 2005 содержат достаточно мало документов в суточном окне, что приводит к небольшому количеству кластеров с размерами более 10 документов, что существенно отличается от характеристик современных потоков новостных сообщений.

Желательно провести аналогичное исследование для более представительных новостных коллекций.

Литература

[1] Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Российский семинар по оценке методов информационного поиска. Труды второго российского семинара РОМИП'2004 (Пущино, 01.10.2004) – СПб: НИИ Химии СПбГУ. – 2004. – С. 62-89.

[2] Амзин А. Новостная интернет-журналистика. – http://kebati.ru/journ/journ.pdf.

[3] Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров // Труды 10-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2008. – Дубна, 2008. – С. 220-229.

[4] Антонов А.В., Баглей С.Г., Мешков В.С. Подход к выявлению подмножеств похожих документов // Труды 10-й Всерос. науч. конф. «Электронные библиотеки: перспективные ме-

тоды и технологии, электронные коллекции» RCDL'2008. – Дубна, 2008. – С. 197-199.
[5] Васильев В.Г., Кривенко М.П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008. – 302 с.
[6] Васильева Л.А. Делаем новости! – Учебное пособие. – М.: Аспект Пресс, 2003.
[7] Воронцов К.В. Лекции по алгоритмам кластеризации и многомерного шкалирования. – www.ccas.ru/voron/download/Clustering.pdf.
[8] Григорян М. Пособие по журналистике. – М.: «Права человека», 2007. – 192 с.
[9] Добров Б.В., Лукашевич Н.В. Автоматическая обработка больших массивов англоязычных текстов // Труды межд. семинара «Когнитивное моделирование», Пущино, 1999. – fccl.ksu.ru/winter.99/cog_model/englproc.pdf.
[10] Лукашевич Н.В., Салий А.Д. Представление знаний в системе автоматической обработки текстов // НТИ. Сер. 2. – 1997. – № 3 – www.cir.ru/docs/ips/publications/1997_nti_thes.pdf.
[11] Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука, 1985.
[12] Кондратьев М.Е. Анализ методов кластеризации новостного потока // Труды 8-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2006. – Суздаль, 2006. – С. 108-114.
[13] Пескова О.В. Разработка метода автоматического формирования рубрикатора полнотекстовых документов. – Дис. … канд. техн. наук. – Москва, МГТУ им. Н. Э. Баумана, 2008.
[14] Amigo E., Gonzalo J., Artiles J. A comparison of extrinsic clustring evaluation metrics on formal constraints // Information Retrieval. – 2009.
[15] Christopher D.M., Prabhakar R., Hinrich S. Introduction to information retrieval. – Cambridge University Press, 2008. – http://nlp.stanford.edu/IR-book/information-retrieval-book.html.
[16] Dezs Z., Almaas E., Lukacs A., Racz B., Szakadat I., Barabasi A.-L. Dynamics of information access on the web// Physical Review. – 2006. – E 73, 066132.
[17] Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // E. Simoudis, J. Han, U.M. Fayyad. Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. – 1996. – P. 226-231.
[18] Jain A.K., Murty M.N., Flynn P.J. Data clustering: a review // ACM Computing Surveys (CSUR) Archive. – 1999. – V. 31, Issue 3. – P. 264-323.
[19] Wu J., Xiong H., Chen J. Adapting the right measures for K-means clustering // Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. – Paris, France, 2009. – P. 877-886.
[20] Lewis D.D., Yang Y., Rose T.G., Li F. RCV1: a new benchmark collection for text categorization research // J. of Machine Learning Research. – 2004. – V. 5. – P. 361-397.
[21] Sanderson M. Reuters test collection // Proc. of the Sixteenth Research Colloquium of the British Computer Society Information Retrieval Specialist Group, Drymen, 1994.
[22] Zhong S., Gosh J. A unified framework for model-based clustering // J. of Machine Learning Research. – 2003. – V. 4. – P. 1001-1037.

Basic line for news clusterization methods evaluation

B. Dobrov, A. Pavlov

The paper is devoted to evaluation of several news clusterization methods – agglomerative, k means, DBSCAN and FOREL. The authors made manual partition for three Wednesdays of data collection ROMIP 2006 (http://romip.ru/en/collections/news-collection. html). All studied methods shown similar results in F1-measure with average value about 72 % with difference from 58 % up to 87 % for different days.