УДК 681.3

А.А. Шумейко, С.Л. Сотник

ИСПОЛЬЗОВАНИЕ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ ДЛЯ АВТОМАТИЧЕСКОЙ РУБРИКАЦИИ ТЕКСТОВ

Статья посвящена построению рубрикатора с использованием агломеративной кластеризации. В качестве критерия близости используется сводный коэффициент обобщенной корреляции в предположении, что случайные величины имеют нормальное распределение и взаимосвязь между ними вычисляется в соответствии с критерием Стьюдента.

Ключевые слова: обработка текстов, кластеризация, автоматическая рубрикация, корреляция слов

Введение

Человеку присуще умение и стремления все раскладывать в определенном порядке, упорядочивая свои знания и представления об окружающем мире. На первом этапе упорядочения тех или других объектов происходит их объединения в качественно однородные группы, то есть группирование. На более высоком уровне развития процесса исследования объектов они делятся на классы на основе общности и расхождении присущих им признаков, которые отражают взаимосвязи между классами, то есть классифицируются. В результате создается система классификации. В классификации отражается суть объектов, которые изучаются на основе общих признаков и отношений между ними. Классификация имеет большое значение в библиотечно-библиографической деятельности. В библиотеках постоянно возникает необходимость распределять на классы (группы) документы на основе их похожести и различия между ними. Классификационные системы, в которых предполагается распределение документов по смыслу, делятся на систематические и предметные. В систематических классификационных системах основой деления является область знания, а в предметных - непосредственно сам предмет. Ведущее место принадлежит систематическим системам классификации. Такие системы служат примером естественной системы класси-

фикации, так как распределение документов в них (систематизация) осуществляется по важным признакам - по областям знаний на основе содержания документов. Эти системы используются для организации систематических каталогов и картотек. Важное место в систематизации занимают рубрики. Рубрикатор воссоздает отраслевое распределение народного хозяйства, распределение по научным дисциплинам, а также по комплексным межотраслевым проблемам (охрана окружающей среды, охрана работы, исследования космического пространства и др.), что отвечает современной тенденции дифференциации и интеграции разных областей знаний.

Переход на использование безбумажных, электронных носителей информации подразумевает использование автоматизированных методов классификации и систематизации поступающих документов, таким образом, все это делает задачу автоматизации методов рубрикации актуальной.

Постановка задачи

Формально рубрикатор представляет собой дерево, каждый последующий уровень которого детализирует область, описанию которой посвящен данный документ. При этом каждый уровень характеризуется одним или несколькими словосочетаниями. Заметим, разбиения этих словосочетаний на множества ключевых слов, нарушает идеологию рубрики, например, сочетания «непрерывная функция» и «непрерывная разливка» определяют совершенно разные рубрики. Таким образом, ставится задача автоматического определения термов (словосочетаний), характеризующих данную группу документов и по наборам полученных термов определение соответствующих рубрик. Решению данной проблемы посвящена данная статья.

Пусть $\Im = \{d_i\}_{i=1}^{n}$ множество документов, определенных набором наиболее характерных слов $d_i = \{w_1^i, w_2^i, ..., w_{m_i}^i\}$, где w_k^i принимает значение из заданного множества W. Задача кластеризации состоит в построении множества $C = \{c_v\}_{v=1}^{N}$ и отображения $F : \Im \to C$ заданного множества на множество кластеров. Кластер содержит объекты из $\Im$ похожие (по заданному критерию) друг на друга $d_i, d_j \in c_v \Rightarrow dis(d_i, d_j) < \varepsilon$, где $dis(a,b)$ мера близости (расстояние) меж-

ду двумя объектами, а ε - максимальное значение порога, формирующее один кластер.

Условно методы кластеризации разбиваются на два класса - иерархические и неиерархические [1, 2]. В неиерархических алгоритмах является наличие условия остановки и количества кластеров. Основой этих алгоритмов является гипотеза о сравнительно небольшом числе скрытых факторов, которые определяют структуру связи между признаками. Иерархические алгоритмы не завязаны на количестве кластеров, эта характеристика определяется по динамике слияния и разделения кластеров, во время построения дерева вложенных кластеров (дендрограммы). В свою очередь, иерархические алгоритмы делятся на агломеративные, которые строятся путем объединения элементов, то есть уменьшением количества кластеров, и дивизимные, основанные на разделении (расщеплении) существующих групп (кластеров). Задаче автоматической рубрикации посвящены работы [4,5], кластеризация текстовых документов на основе неиерархических алгоритмов подробно рассматривалась в работах [6,7,8].

Основные результаты

Применим идею агломеративной кластеризации к формированию устойчивых словосочетаний, характеризующих данную группу документов.

Каждому слову $w_k \in W$ ставится в соответствие вектор $M_k = \{m_1^k, m_2^k, ..., m_n^k\}$, где m_i^k частота встречи этого слова в документе d_i. Построим дерево, на нижнем уровне которого будут пары слов (w_ν, w_μ) (при этом $(w_\mu, w_\nu) = (w_\nu, w_\mu)$), характеризуемые коэффициентом обобщенной корреляции

$$r_{\nu,\mu} = \frac{\sum_{i=1}^{n}\left(m_i^\mu - \overline{m}^\mu\right)\left(m_i^\nu - \overline{m}^\nu\right)\omega\left(m_i^\mu, m_i^\nu, \overline{m}^\mu, \overline{m}^\nu\right)}{\sqrt{\sum_{i=1}^{n}\left(m_i^\mu - \overline{m}^\mu\right)^2}\sqrt{\sum_{i=1}^{n}\left(m_i^\nu - \overline{m}^\nu\right)^2}}, \quad \overline{m}^\nu = \frac{1}{n}\sum_{i=1}^{n} m_i^\nu, \overline{m}^\mu = \frac{1}{n}\sum_{i=1}^{n} m_i^\mu.$$

Здесь $\omega\left(m_i^\mu, m_i^\nu, \overline{m}^\mu, \overline{m}^\nu\right)$ представляет собой весовую функцию общего вида, которая может зависеть как от конкретных элементов векторов частот, так и от каких-то их общих характеристик (например, норм векторов).

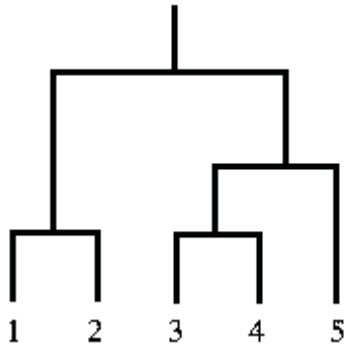

Рисунок 1 - Дерево агломеративной кластеризации

Среди всех элементов нижнего уровня выберем элемент с наибольшим коэффициентом обобщенной корреляции. Пусть это будет элемент (w_ν, w_μ). Для всех (w_μ, w_ζ) $(\zeta \neq \nu)$ построим обобщенную корреляционную матрицу

$$R_{\nu,\mu,\zeta} = \begin{pmatrix} r_{\nu,\nu} & r_{\nu,\mu} & r_{\nu,\zeta} \\ r_{\mu,\nu} & r_{\mu,\mu} & r_{\mu,\zeta} \\ r_{\zeta,\nu} & r_{\zeta,\mu} & r_{\zeta,\zeta} \end{pmatrix}.$$

Через

$$\Lambda = R^{-1}_{\nu,\mu,\zeta} = \begin{pmatrix} \lambda_{\nu,\nu} & \lambda_{\nu,\mu} & \lambda_{\nu,\zeta} \\ \lambda_{\mu,\nu} & \lambda_{\mu,\mu} & \lambda_{\mu,\zeta} \\ \lambda_{\zeta,\nu} & \lambda_{\zeta,\mu} & \lambda_{\zeta,\zeta} \end{pmatrix}$$

Обозначим матрицу, обратную к ней. В качестве меры корреляции между w_ν и w_μ, w_ζ, будем использовать сводный коэффициент обобщенной корреляции

$$r_{\nu,\mu,\zeta} = \sqrt{1 - \frac{1}{r_{\nu,\nu}\lambda_{\nu,\nu}}}.$$

Среди всех $r_{\nu,\mu,\zeta}$ $(\zeta \neq \nu,\mu)$ выберем коэффициент с наибольшим значением. Предполагая, что случайные величины имеют нормальное распределение, взаимосвязь между случайными величинами вычисляется в соответствии с критерием Стьюдента

$$\xi_{\nu,\mu,\zeta} = 0.5 \ln\left(\frac{(1 + r_{\nu,\mu})(1 - r_{\nu,\mu,\zeta})}{(1 - r_{\nu,\mu})(1 + r_{\nu,\mu,\zeta})}\right)\sqrt{\frac{n-3}{2}}.$$

И в случае, если для заданного порога ε выполняется условие $\xi_{\nu,\mu,\zeta} > \varepsilon$, то из элементов нижнего уровня удалим (w_ν, w_μ) и (w_μ, w_ζ).

 ISSN 1562-9945

Полученная тройка слов w_ν и w_μ, w_ζ формирует терм, используемый в дальнейшем для определения подходящей рубрики. Если же это условие не выполняется, то выбираем пару (w_i, w_j) с коэффициентом обобщенной корреляции предшествующим $r_{\nu,\mu}$.

Далее этот процесс повторяем, в результате получаем набор термов, состоящих из неопределенного числа слов, определяемых только порогом ε сводного коэффициента корреляции.

Для определения корневого уровня дерева рубрикации, для всех термов (словосочетаний) корневого уровня рубрик, нужно найти соответствующий терм из построенного дерева с самым большим сводным коэффициентом обобщенной корреляции. Для определения следующего уровня вложения выбирается множество $\Re$ термов содержащих хоть одно слово из терма соответствующего корневому уровню. Для соответствующих рубрик первого уровня вложения находим соответствующий терм из $\Re$ с самым большим сводным коэффициентом корреляции. Далее этот процесс повторяется требуемое число раз, либо, пока не закончится дерево, определяемое значением порога ε.

Заметим, что при условии $\omega\left(m_i^\mu, m_i^\nu, \overline{m}^\mu, \overline{m}^\nu\right) \equiv 1$ коэффициент обобщенной корреляции совпадает с коэффициентом корреляции Пирсона. Результаты, полученные в этом случае уступают по своей эффективности с при применением весовых функций вида

$$\omega\left(m_i^\mu, m_i^\nu, \overline{m}^\mu, \overline{m}^\nu\right) = \sqrt{\sum_{i=1}^{n}\left(m_i^\mu - \overline{m}^\mu\right)^2}\sqrt{\sum_{i=1}^{n}\left(m_i^\nu - \overline{m}^\nu\right)^2}$$

и

$$\omega\left(m_i^\mu, m_i^\nu, \overline{m}^\mu, \overline{m}^\nu\right) = 2^n\left(\frac{\sqrt{\sum_{i=1}^{n}\left(m_i^\mu - \overline{m}^\mu\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(m_i^\nu - \overline{m}^\nu\right)^2} + \varepsilon} + \frac{\sqrt{\sum_{i=1}^{n}\left(m_i^\nu - \overline{m}^\nu\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(m_i^\mu - \overline{m}^\mu\right)^2} + \varepsilon} + \varepsilon\right)^{-n}$$

Данный вид весовой функции позволил выделять взаимосвязанные термины при сравнительно небольшом (по сравнению со многими другими весовыми функциями) шуме.

Здесь ε - небольшое число, например 10^{-6}, чтобы избежать деления на ноль. В этом случае наибольший вес (единица) появлялся у тех терминов, которые встречаются с одинаковой частотой. Парамет-

ром n можно регулировать крутизну наклона характеристики. Такая весовая функция позволяет выделять слова, связанные связями типа "синоним", "антоним". В результате получается достаточно много шумовых данных, однако выделив связи типа «общее-частное», можно отфильтровать только те связанные термины, которые также имеют общих родителей.

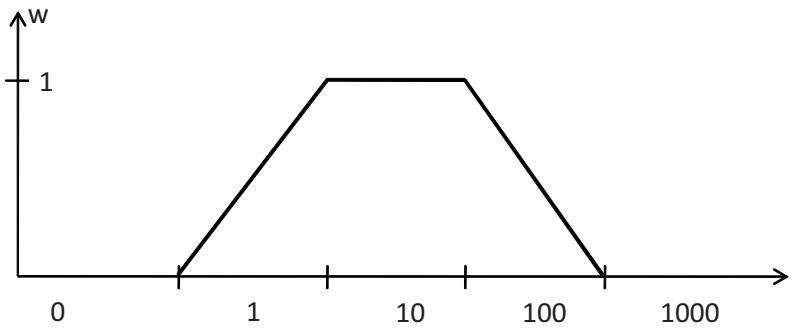

Рисунок 2 - Пример используемой весовой функции

В качестве еще одного примера рассмотрим кусочно-линейную функцию, изображенную на рисунке 2. В этом случае аргументом выступает отношение норм векторов частот. Данный вид весовой функции позволяет выделить связи типа «общее-частное» (голонимы и гиперонимы), поскольку для них характерно, что общий термин встречается на порядок-два чаще, чем специализированный.

Выводы

Результаты работы были внедрены в поисковике HulbeeSearcher (www.hulbee.com) и показали высокую эффективность в проведении автоматической рубрикации документов.

ЛИТЕРАТУРА

1. Bradley P. Scaling Clustering Algorithms to Large Databases / P.Bradley, U.Fayyad, C.Reina // Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining .- AAAI Press: Menlo Park (Calif.), 1998.
2. Zhang T. An Efficient Data Clustering Method for Large Databases / T. Zhang, R.Ramakrishnan, M. Birch Livny // Proc. ACM SIGMOD Int'l Conf. Management of Data .- ACM Press: New York, 1996.
3. Киселев М.В. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики / М.В. Киселев, В.С. Пивоваров, М.М. Шмулевич // Междунар. сб. науч. раб.: Интернет-математика 2005: автоматическая обработка веб-данных. – М.: Яндекс, 2005.

4. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System / Chen Hsinchun, D. Ng Tobun, Joanne Martinez, Bruce R. Schatz // Journal of the American Society for Information Science .- 1997 .- 48(1) .- P. 7-31.
5. Chen H. An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Branch-and-Bound Search vs. Connectionist Hopfield Net Activation / Chen Hsinchun, D. Ng Tobun // Journal of the American Society for Information Science .- 1995 .- 46(5) .- P. 348-369.
6. Шумейко А.А. Использование генетических алгоритмов в задачах классификации текстов / А.А. Шумейко, С.Л. Сотник, М.В. Лысак // Системные технологи .- Днепропетровск, 2009 .- 1(60) .- С. 125-138.
7. Шумейко А.А. Итерационный метод построения векторного классификатора / А.А. Шумейко, С.Л. Сотник // Математичне моделювання .- ДДТУ, 2009 .- №1(20) .- С. 7-11.
8. Shumeyko A. Using Genetic Algorithms for Texts Classification Problems / A. Shumeyko, S. Sotnik // Anale.Seria Informatica .- 2009 .- Vol.VII facs.1 .- P. 325-340.