КОМПЬЮТЕРНЫЕ НАУКИ

УДК004.6

КЛАСТЕРИЗАЦИЯ СЛАБОСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ

ГВОЗДИНСКИЙ А.Н., ГУБИН В.А., ЮРДИГА Л. А.

Предлагается метод, позволяющий осуществить кластеризацию слабоструктурированных текстовых документов. Кластеризация основана на учете вхождения в документы текстовых строк, классифицированных как атрибуты данных. В качестве критерия оптимальности используется функционал, достижение которым максимального значения означает завершение процесса кластеризации. Обосновывается выбор метода и параметров кластеризации.

1. Введение

Постановка задачи. В данной работе решается задача кластеризации слабоструктурированных текстовых документов. В статье [1] представлены их основные признаки и свойства и приведены примеры такого рода документов. К ним можно отнести анкеты, страховые формы, налоговые декларации, счета, транспортные накладные, контракты, технические параметры изделия, прайс-листы, типовые договора, карточки учета, результаты спортивных матчей.

Предлагаемый здесь метод кластеризации ставит своей основной задачей разбить исходное множество документов на подмножества, каждое из которых представляет подобные документы. При этом предполагается не семантическое подобие на уровне тематики документов, а типовое подобие документов, при котором тип документа определяют входящие в него текстовые фрагменты, являющиеся атрибутами данных.

Таким образом, особенностью рассматриваемого процесса кластеризации является то, что предлагаемый метод учитывает вхождение (или отсутствие такового) в документ не слов (термов), а обособленных текстовых фрагментов документа, являющихся атрибутами данных, которые и определяют принадлежность документа к тому или иному типу. Другими словами, автор исходит из того, что каждый тип документов характеризуется множеством используемых в документах этого типа атрибутов.

Анализ последних достижений и публикаций. Одной из проблем, возникающих в процессе кластеризации текстовых документов, является большая размерность их векторной модели. Частично острота проблемы снимается применением тех или иных методов предварительной фильтрации, используемых при написании документов слов. Методы, решающие эту проблему более фундаментально, основаны на учете специфики содержимого текстовых документов и решаемых на множестве текстовых документов задач.

Так, в работах [2,3] были предложены и исследованы подходы, основанные на замене термов документа концептами. Отмечается, что добавление концептов в векторные представления документов имеет два преимущества: первое – устранение синонимов, второе - возможность, используя онтологии, выводить более общие концепты. Таким образом, для сокращения размерности предлагается преобразовывать вектор термов документа в вектор концептов документа.

В работе [4] представлены два подхода, повышающих эффективность процесса кластеризации: кластеризация на основе часто повторяемых последовательностей слов и кластеризация на основе часто повторяемых последовательностей значений слов. Отмечается, что ключевой особенностью этих алгоритмов является то, что они относятся к тексту документа как к последовательности слов (значений слов), а не как к набору слов.

В данной работе предлагается подход, при котором документы, подвергаемые кластерному анализу, представляются в виде набора объектов, каждый из которых представляет текстовый фрагмент, являющийся атрибутом данных.

Цели и задачи исследования. Процесс кластеризации, рассматриваемый в данной работе, трактуется как дискретная оптимизация, при которой необходимо каждому документу поставить в соответствие номер кластера так, чтобы достиг своего экстремального значения некоторый критерий оптимальности.

Соответственно, целью данной работы является формулирование и формальное выражение критерия оптимальности и уточнение параметров кластеризации, при которых, по мере объединения документов в кластеры, будет улучшаться значение данного критерия.

Уточнения и обоснования также требует выбор метода кластеризации, выбор меры расстояния (меры подобия) между документами и между кластерами, выбор метода объединения.

2. Построение критерия оптимальности процесса кластеризации

Пусть имеется некоторое множество слабоструктурированных текстовых документов $\Omega = \{D_1, D_2, ..., D_N\}$. Предположим, что получено некоторое разбиение множества Ω на кластеры $K_j, j = \overline{1, N_k}$, где N_k – количество кластеров. Исходим из того, что каждый документ входит только в один из кластеров. В этом случае множество документов Ω можно представить следующим образом:

$$\Omega = K = \{K_1, K_2, ..., K_{N_k}\}.$$

Пусть ψ – текстовая строка документа, являющаяся атрибутом данных. Будем говорить, что некоторый атрибут ψ принадлежит некоторому кластеру K_j: $\psi \in K_j$, если данный атрибут принадлежит хотя бы одному документу из кластера K_j. Обозначим через P_{ψ,K_j} частоту встречаемости атрибута ψ в документах кластера K_j при условии, что $\psi \in K_j$, т.е.

$$P_{\psi,K_j} = \frac{|D_i, \psi \in D_i, D_i K_j|}{|D_i, D_i K_j|}.$$

При этом знаменатель равен общему количеству документов в кластере K_j, а числитель – количеству документов в кластере K_j, содержащих атрибут ψ.

Обозначим через $P_{\psi,K/K_j}$ частоту встречаемости атрибута ψ в документах вне кластера K_j при условии, что $\psi \in K_j$, т.е.

$$P_{\psi,K/K_j} = \frac{|D_i, \psi \in D_i, D_i K/K_j|}{|D_i, D_i K/K_j|}. \quad (1)$$

При этом знаменатель равен общему количеству документов вне кластера K_j, а числитель – количеству документов вне кластера K_j, содержащих атрибут ψ.

И, наконец, обозначим через $\overline{P}_{\psi,K/K_j}$ частоту не встречаемости атрибута ψ в документах вне кластера K_j при условии, что $\psi \in K_j$, т.е.

$$\overline{P}_{\psi,K/K_j} = \frac{|D_i, \psi \notin D_i, D_i K/K_j|}{|D_i, D_i K/K_j|}. \quad (2)$$

Заметим, что

$$\overline{P}_{\psi,K/K_j} = 1 - P_{\psi,K/K_j}, \quad (3)$$

так как частоты (1) и (2) соответствуют оценкам вероятностей противоположных событий.

Ожидается, что в каждое множество K_i входят документы, относящиеся к одному типу. Поскольку тип документа во многом определяют входящие в него атрибуты данных, логично предположить, что в документы из одного и того же кластера входят преимущественно одни и те же атрибуты. И так же логично предположить, что если документы не принадлежат одному и тому же кластеру, в них входят преимущественно различные атрибуты. Можно сказать и по-другому: если некоторый атрибут данных принадлежит некоторому документу, то из этого следует, что вероятность того, что этот же атрибут содержится и в других документах этого же кластера должна быть существенно больше вероятности того, что этот же атрибут встречается в документах, не принадлежащих данному кластеру.

Другими словами, для большинства пар (ψ_k, K_j), где $\psi_k \in K_j$, при качественном разбиении множества документов Ω на кластеры мы должны иметь следующее:

$$P_{\psi,K_j} >> P_{\psi,K/K_j}.$$

В идеальном варианте, когда в документах из одного кластера содержатся одни и те же наборы атрибутов и при этом их нет в документах из других кластеров, будем иметь следующее:

$$P_{\psi,K_j} = 1, P_{\psi,K/K_j} = 0.$$

Таким образом, хорошим будет такое разбиение на кластеры, которое для всех пар (ψ_k, K_j), $\psi_k \in K_j$ максимизирует сумму частот P_{ψ_k,K_j}:

$$F_1(K) = \sum_{(\psi_k,K_j),\psi_k \in K_j} \frac{|K_j|}{|K|} p_k P_{\psi_k,K_j} \to \max \quad (4)$$

и одновременно минимизирует сумму частот $P_{\psi_k,K/K_j}$:

$$F_2(K) = \sum_{(\psi_k,K_j),\psi_k \in K_j} \frac{|K/K_j|}{|K|} p_k P_{\psi_k,K/K_j} \to \min. \quad (5)$$

Коэффициенты $\frac{|K_j|}{|K|}$ в (4) и $\frac{|K/K_j|}{|K|}$ в (5) призваны сделать сопоставимыми масштабы выражений (4) и (5), а p_k – весовой коэффициент, равен вероятности того, что ψ является атрибутом данных.

С учетом (3) (5) можно переписать следующим образом:

$$F_2(K) = \sum_{(\psi_k,K_j),\psi_k \in K_j} \frac{|K/K_j|}{|K|} p_k (1 - \overline{P}_{\psi_k,K/K_j}) \to \max \quad (6)$$

Свернув критерии (5) и (6) в один, получим, что хорошим будет такое разбиение K множества слабоструктурированных текстовых документов на кластеры, при котором достигает своего максимума значение следующего функционала:

$$F(K) = \lambda_1 F_1 + \lambda_2 F_2 \to \max, \quad (7)$$

где λ_1, λ_2 – коэффициенты пропорциональности, устанавливающие баланс взаимного влияния сумм частот (4) и (6), входящих в (7); λ_1, λ_2 могут принимать любые неотрицательные значения. При этом важно значение их отношения.

3. Выбор метода и параметров кластеризации

В качестве основы целесообразно выбрать иерархический метод кластерного анализа, так как, в общем случае, предварительно не известно количество типов документов, к которым можно отнести документы из множества Ω, и, таким образом, нет возможности на начальном этапе выдвинуть гипотезу о количестве кластеров.

Для осуществления процесса кластеризации необходимо определиться с мерой расстояния между документами и между кластерами.

Чем больше будет атрибутов, общих для пары документов, объединяемых в кластер, тем большее приращение получит значение функционала (7). Таким образом, в качестве меры расстояния $r_{ij}(D_i, D_j)$ между документами D_i и D_j, необходимо выбрать следующее выражение:

$$r_{ij}(D_i, D_j) = 1 - \frac{\left|D_i \cap D_j\right|_\psi}{\left|D_i \cup D_j\right|_\psi}, \qquad (8)$$

где $\left|D_i \cap D_j\right|_\psi$ – количество атрибутов, содержащихся в пересечении документов D_i и D_j, а $\left|D_i \cup D_j\right|_\psi$ – количество атрибутов, содержащихся в объединении документов D_i и D_j.

Наиболее популярными методами объединения кластеров являются следующие: метод ближнего соседа или одиночная связь, метод наиболее удаленных соседей или полная связь, метод попарного среднего.

Далее будет дана трактовка и выражение расстояния $R_{ij}(K_i, K_j)$ между кластерами K_i и K_j для каждого метода объединения кластеров.

1. *Метод ближнего соседа или одиночная связь.* Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими документами (ближайшими соседями) в различных кластерах. В этом случае расстояние между кластерами можно выразить следующим образом:

$$r_{i'j'}(D_{i'}, D_{j'}).$$

2. *Метод наиболее удаленных соседей или полная связь.* Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. “наиболее удаленными соседями”). В этом случае расстояние между кластерами можно выразить следующим образом:

$$r_{i'j'}(D_{i'}, D_{j'}).$$

3. *Метод попарного среднего.* Здесь в качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. В этом случае расстояние между кластерами можно выразить следующим образом:

$$R_{ij}(K_i, K_j) = \frac{\sum_{(D_{i'}, D_{j'})} r_{i'j'}(D_{i'}, D_{j'})}{|(D_{i'}, D_{j'})|}, D_{i'} K_i, D_{j'} K_j.$$

где $|(D_{i'}, D_{j'})|$ – количество всевозможных пар, образованных документами $D_{i'} \in K_i$ и $D_{j'} \in K_j$.

Необходимо заметить, что если в каждом из кластеров K_i, K_j по одному документу D_i, D_j, то независимо от выбранного метода объединения $R_{ij}(K_i, K_j) = r_{ij}(D_i, D_j)$.

Возникает вопрос, какой из этих методов объединения выбрать? Известно, что при выборе метода одиночной связи могут получиться цепочечные кластеры. В нашем случае это может привести к тому, что в один кластер могут быть объединены документы родственных, но разных типов, если при этом эти типы содержат общие атрибуты данных.

При выборе метода полной связи есть риск попадания в разные кластеры документов, объективно принадлежащих к одному типу. Это возможно в тех случаях, когда некоторый тип документов допускает заметные отклонения от документа к документу с точки зрения используемых в них наборов атрибутов данных.

В свете изложенного некоторым компромиссом выглядит метод попарного среднего. Поэтому его наиболее целесообразно использовать при разбиении на кластеры множества слабоструктурированных документов. При этом, варьируя значения параметров λ_1, λ_2, можно сделать этот процесс более либеральным, как это происходит при методе одиночной связи, либо более избирательным, как это происходит при методе полной связи.

4. Описание процесса кластеризации

Процесс кластеризации протекает следующим образом. На начальном этапе строится матрица $R = \| r_{ij} \|_{NN}$ расстояний между парами документов, N – количество анализируемых документов. Для общности изложения будем считать, что на начальном этапе каждый документ образует отдельный кластер. Далее:

1. Предпринимается попытка нахождения наиболее близких кластеров для последующего их объединения, т.е. объединяются в один те кластеры, для которых $R_{ij}(K_i, K_j)$ достигает своего минимального значения.

2. Проверяется значение функционала (7).

3. Если это значение не уменьшилось, то данное объединение засчитывается, матрица R пересчитывается и осуществляется переход к пункту 1.

4. Если же значение функционала (7) уменьшилось, то данное объединение отменяется и предпринимается попытка объединения другой пары наиболее близких

между собой кластеров. Осуществляется переход к пункту 2.

5. Если такая пара кластеров отсутствует, то процесс кластеризации завершается.

5. Выводы

Сформулирован и формально выражен критерий оптимальности и уточнены параметры кластеризации. Обоснован выбор метода кластеризации, выбор меры расстояния (меры подобия) между документами и между кластерами, выбор метода объединения.

Научная новизна: получили дальнейшее развитие иерархические агломеративные методы кластерного анализа текстовых документов. Предложен метод кластерного анализа, основанный на учете вхождения в документы текстовых фрагментов, являющихся атрибутами данных, что дает возможность разбивать исходное множество документов на подмножества, каждое из которых представляет документы одного типа.

Практическая значимость: использование разработанного метода кластерного анализа позволяет осуществлять кластеризацию текстовых документов, в основе которой лежит типовое сходство документов.

Направления дальнейших исследований: создаются предпосылки автоматизации построения обучающих выборок документов при организации извлечения из документов данных конкретных типов.

Литература: 1. *Губин В.А.* Слабоструктурированные текстовые документы как источники данных // Бионика интеллекта. Х.: ХНУРЕ, 2010. №3(74). С. 109–111. **2.** *Andreas Hotho, Steffen Staab, Gerd Stumme.* Ontologies Improve Text Document Clustering // Proc. of the 2003 IEEE International Conference on Data Mining, Poster. Melbourne, Florida, IEEE Computer Society, November 19-22, 2003. P. 541-544. **3.** *Shady Shehata, Fakhri Karray, Mohamed Kamel.* Enhancing Text Clustering Using Concept-based Mining Model // Proceedings of the Sixth International Conference on Data Mining, IEEE Computer Society Washington, DC, USA, 2006. P. 1043-1048. **4.** *Yanjun Li, Soon M. Chung. Text* Document Clustering Based on Frequent Word Sequences // Proceedings of the 14th ACM international conference on Information and knowledge management. New York, USA, ACM Press, 2005. P. 293-294.

Поступила в редколлегию 02.03.2011

Рецензент: д-р техн. наук, проф. Куземин А.Я.

Гвоздинский Анатолий Николаевич, канд. техн. наук, профессор кафедры искусственного интеллекта ХНУРЭ. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 70-21-337.

Губин Вадим Александрович, ст. преподаватель кафедры искусственного интеллекта ХНУРЭ. Адрес: Украина, 61166, Харьков, пр..Ленина, 14, тел. 70-21-337.

Юрдига Любовь Антоновна, студентка гр. КН-07-5 ХНУРЭ. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 70-21-337.