Вычислительные технологии, 2020, том 25, № 3, с. 99–110. © ИВТ СО РАН, 2020 ISSN 1560-7534
Computational Technologies, 2020, vol. 25, no. 3, pp. 99–110. © ICT SB RAS, 2020 eISSN 2313-691X

DOI:10.25743/ICT.2020.25.3.011

Кластеризация документов на основе семантической матрицы связей для концептуального индексирования

Т. В. Авдеенко†, Ю. А. Мезенцев
Новосибирский государственный технический университет, Новосибирск, Россия
†Контактный автор: Авдеенко Татьяна В., e-mail: avdeenko@corp.nstu.ru

Поступила 10 марта 2020 г., доработана 20 марта 2020 г., принята в печать 14 апреля 2020 г.

Рассматривается проблема кластеризации — автоматического разбиения коллекции документов на группы, близкие по тематике. Предложен новый подход к концептуальному индексированию документов путем перехода от набора ключевых терминов к взвешенному множеству концептов некоторой иерархической модели знаний. Получаемая в результате применения данного метода семантическая матрица связей документов с концептами онтологии может быть использована в качестве матрицы данных для решения задачи кластерного анализа. Оригинальный подход к кластеризации сводится к формализации в виде NP-трудной задачи смешанного программирования, декомпозиции и поэтапному решению, снижающему ее трудоемкость.

Ключевые слова: кластеризация документов, концептуальное индексирование, таксономия, онтология, смешанное целочисленное программирование, NP-трудная задача.

Цитирование: Авдеенко Т.В., Мезенцев Ю.А. Кластеризация документов на основе семантической матрицы связей для концептуального индексирования. Вычислительные технологии. 2020; 25(3):99–110.

Введение

Кластеризация документов (текстов) — это сфера исследований, которая является подобластью более широкой области кластеризации данных, имеющей много общего с такими областями, как извлечение информации (Information Retrieval), обработка естественного языка (Natural Language Processing), машинное обучение (Machine Learning). Процесс кластеризации направлен на то, чтобы обнаружить естественные кластеры документов и соответственно сгруппировать их по смысловым признакам (темам) в сборнике (базе данных) документов для последующего более эффективного их поиска (извлечения).

В области искусственного интеллекта задача кластеризации данных относится к тематике неконтролируемого машинного обучения (обучение без учителя), когда распределение объектов по классам заранее неизвестно, в отличие от задачи классификации, которая относится к тематике контролируемого машинного обучения (обучение с учителем), когда принадлежность объектов к классам задается в обучающей выборке. Кластеризация документов определена в [1] как оптимизационный процесс, который направлен на разбиение коллекции документов таким образом, чтобы документы внутри

каждого кластера были наиболее сходны (компактность кластеров), но при этом сформированные кластеры были бы как можно более различны по смысловому содержанию (различимость кластеров).

Процесс кластеризации документов может быть разбит на два основополагающих этапа. Первый этап связан с предобработкой документов и имеет целью их преобразование в форму, подходящую для проведения анализа данных. Второй этап (собственно алгоритм кластеризации) нацелен непосредственно на анализ подготовленных на первом этапе данных и разделение документов на кластеры. Несмотря на важность алгоритма кластеризации, нельзя недооценивать первый этап процесса кластеризации, который столь же важен, как и, собственно, сам оптимизационный алгоритм [2].

В качестве первого этапа предобработки документов традиционно используется индексирование — замена коллекции документов последовательностью индексных терминов (ключевых слов). Индексные термины могут быть получены непосредственно из документа или сгенерированы опосредованно с использованием более сложных алгоритмов. Методы индексирования различаются, во-первых, способами определения индексных терминов, во-вторых, методами вычисления важности термина относительно рассматриваемого документа [3]. Большинство методов индексирования основаны на традиционных подходах с использованием набора терминов (ключевых слов) и метода вычисления частоты встречаемости термина (TF — Term Frequency). В процессе применения данных подходов определяется множество индексных терминов и вычисляются их веса на основе частоты повторения термина в документе. В результате каждый документ D индексируется вектором **TF**. Чем выше вес термина в векторе **TF**, тем выше его значимость в документе.

Несмотря на широкое применение, методы, использующие данный подход к индексированию, страдают от следующих недостатков:

— не дают возможности дифференцировать семантическую важность каждого термина в документе, так как присваивают веса на основе частоты встречаемости термина, не учитывая семантическую важность слов в документе;
— не учитывают синонимы, многозначные слова и т. д., поэтому документ и центроидные векторы могут содержать различные индексные термины, которые являются, например, синонимами [4];
— скорее всего, ведут к векторам очень большой размерности (полнотекстовые индексы).

Для преодоления указанных недостатков классических методов индексирования в литературе предлагаются подходы, различными способами добавляющие семантический компонент к техникам индексирования. Так, в работе [5] описан метод, который использует алгоритм Ripper, который был адаптирован для преодоления проблемы многомерности и многозначности для решения задачи классификации документов. Авторы используют WordNet для извлечения множества синонимов и гиперонимов для каждого существительного и глагола в корпусе и используют эту информацию для преодоления проблемы многозначности. Кроме того, в этой работе впервые используется параметр глубины иерархии (h) для контроля уровня обобщения множества подчиненных концептов. В работе [6] применяется EDR (электронный словарь) для задания связей концептов с ключевыми словами документа. Однако этот подход, как и предыдущий, полностью не решает проблему устранения двусмысленности. Для русскоязычных текстов в диссертации [7] предложен метод концептуального индексирования документов для информационно-поисковой системы, базирующийся на знаниях, описанных в пред-

метно-ориентированной базе знаний, и построенный на тематическом представлении документов. Наиболее продвинутые результаты предлагаются в работе [8], где применен метод концептуального индексирования, заключающийся в попарном сравнении ключевых слов документа с целью поиска наиболее подходящего концепта (смысла), отвечающего наименьшему пути от одного слова к другому на графе, соответствующем таксономии знаний. В качестве таксономии знаний в данной статье используется WordNet. В результате применения такого подхода для каждого документа, проиндексированного ключевыми словами по классической методологии, авторы получают вектор "смыслов" (концептов онтологии WordNet) и весов, определяющих, насколько данный концепт соответствует документу.

В настоящей работе предложен подход дальнейшего усовершенствования процедуры концептуального индексирования документов на основе модели знаний в виде таксономии концептов и представления информации в виде, удобном для последующей кластеризации и извлечения близких по смыслу документов. В качестве таксономии может быть использована иерархическая основа любой предметной онтологии. Для реализации подхода изначально предполагается, что каждый документ имеет взвешенную связь с некоторым множеством (не обязательно одним) концептов вышеуказанной таксономии. Это может быть сделано как вручную, экспертом в соответствующей предметной области, так и автоматически, например, с использованием подхода предобработки документов, описанного в [8]. Предложено построение семантической матрицы связей документов с терминальными (конечными) концептами таксономии, которая может быть использовала как матрица данных для последующего применения методов анализа данных, в частности для решения задачи кластеризации документов.

Для выполнения кластеризации документов в работе использована формальная постановка задачи в виде труднорешаемой задачи смешанного целочисленного программирования. Применение данного подхода (см., например, [9–11]) обосновывается получением теоретически наилучшего (в смысле заданного критерия) решения задачи кластеризации. Это подтверждается экспериментально в сравнении с применением традиционных инструментов кластеризации (например, метода к-центров — см. [12] и нескольких методов машинного обучения, в том числе с применением нейронных сетей Кохонена [13] и SOINN [14]).

1. Метод построения семантической матрицы связей документов с терминальными концептами на основе концептуального индексирования

Предположим, что коллекцию документов необходимо сгруппировать, т. е. разделить на кластеры по тематике. Каждый документ состоит из множества ключевых слов, определяющих его содержание. Применим некоторую иерархическую модель знаний (таксономию или онтологию, например WordNet) той предметной области, к которой относятся документы, подлежащие кластеризации. Каждому документу поставлено в соответствие необходимое число концептов (понятий, смыслов) предметной области, описываемой вышеуказанной иерархической моделью знаний. Это может быть сделано как в ручном режиме (например, экспертами предметной области), так и автоматически. Метод концептуального индексирования [8] позволяет автоматически, на основе множества ключевых слов, определить взвешенное множество концептов, описываю-

щих отдельный документ, и представить его в виде

$$D = \big\{(C_1, \nu_1), (C_2, \nu_2), \ldots, (C_I, \nu_I)\big\}.$$

В этом уравнении использовано обозначение имени концепта C_i и его весового значения ν_i, $0 \le \nu_i \le 1$, $\sum\limits_{i=1}^{I} \nu_i = 1$, устанавливающего силу связи между документом D и соответствующим концептом C_i; I — число связей документа D с концептуальной иерархической моделью. Чем больше весовой коэффициент ν_i, тем ближе смысловое значение рассматриваемого документа соответствующему концепту таксономии.

Выбранный для индексирования документа концепт C_i может быть либо терминальным концептом таксономии (не имеющим подчиненных дочерних концептов), либо нетерминальным (промежуточным). Необходимость установления связи с нетерминальным (более общим) концептом может быть вызвана тем, что в данном документе описывается более общий аспект какого-либо понятия.

Подчеркнем возможность установления связи документа с несколькими концептами таксономии. Это расширяет выразительные возможности предлагаемого подхода, так как предполагает возможность описания междисциплинарной проблемы (документа) на стыке нескольких областей знаний.

В дальнейшем будем различать терминальные и нетерминальные концепты таксономии знаний. Суть предлагаемого подхода заключается в том, что все связи, устанавливаемые между документами и концептами таксономии (не обязательно терминальными), опускаются на уровень терминальных концептов. Соответственно вес ν_i распределяется между терминальными концептами, являющимися потомками концепта C_i. Пусть имеем J терминальных (конечных) концептов таксономии. Для каждого термина kw_j необходимо найти соотношение для вычисления веса w_j, $j = \overline{1, J}$, $\sum\limits_{j=1}^{J} w_j = 1$, на основе начальных весов ν_i, связывающих документ с таксономией, а также с учетом структуры таксономии.

Процедура вычисления весов w_j для терминальных концептов kw_j, $j = \overline{1, J}$, может быть построена следующим образом. Предположим, что документ D связан с концептами $C_1, C_2, \ldots, C_I$ таксономии.

1. Присваиваем начальные значения весам терминальных концептов $w_j = 0$ для всех $\forall\ j = \overline{1, J}$.

2. Выполняем цикл по всем концептам C_i, $i = \overline{1, I}$, связанным с документом:

— если C_i является терминальным концептом ($kw_j = C_i = C_i^{(0)}$), то

$$w_j = \nu_i;$$

— если C_i — нетерминальный концепт, т. е. терминальный концепт kw_j является потомком уровня L промежуточного концепта C_i, $kw_j = C_i^{(L)}$, то

$$w_j = \nu_i \prod_{l=1}^{L} \nu_i^{(l)},$$

где $\nu_i^{(l)}$ есть вес иерархического отношения $R_{ISA}(C_i^{(l-1)}, C_i^{(l)})$ между концептом-родителем $C_i^{(l-1)}$ и концептом-ребенком $C_i^{(l)}$ на пути от концепта $C_i^{(0)}$, связанного с документом D,

к терминальному концепту $C_i^{(L)} = kw_j$. Если в таксономии отсутствуют веса иерархических отношений, то можно положить веса связей, выходящих от одного концепта-родителя, одинаковыми и равными $1/p$, где p — число дочерних концептов.

В результате применения вышеописанного алгоритма к каждому документу из анализируемой коллекции получаем семантическую матрицу связей, число строк которой равно числу документов (объектов в таблице объект — свойство, используемой для анализа данных), число столбцов — количеству терминальных концептов онтологии. Элементы данной таблицы являются числовыми (значения изменяются от 0 до 1), сумма элементов в строке равна 1 согласно процедуре вычисления. К таблице возможно непосредственное применение алгоритмов анализа данных для последующей структуризации пространства документов либо для структуризации пространства терминальных концептов с целью возможного перепроектирования таксономии.

В следующем разделе опишем непосредственно процедуру кластеризации документов с использованием семантической матрицы в качестве таблицы данных.

2. Постановки задачи оптимальной m-кластеризации

Чаще всего используемой мерой близости (расстоянием в многомерном пространстве) является сумма евклидовых расстояний между всеми парами объектов внутри кластера. В вычислительных экспериментах для измерения близости объектов воспользуемся этой метрикой, учитывая возможность ее замены любой другой.

Представим формальную постановку задачи оптимальной m-кластеризации (разбиения на m кластеров) для общего случая. Введем обозначения: $i,\ j = \overline{1,n}$ — номера объектов; $l,\ k = \overline{1,m}$ — номера кластеров; $c_{i,j}$ — расстояние между объектами i и j, $i,\ j = \overline{1,n}$; $c_{i,j}^k$ — расстояние между объектами i и j в кластере k.

Определим переменные y_i^k (идентифицирующие принадлежность объектов $i,\ j = \overline{1,n}$ кластеру $k,\ k = \overline{1,m}$). Определим также зависимые переменные $x_{i,j}^k = y_i^k y_j^k$. Тогда задача кластеризации состоит в определении булевых переменных y_i^k и $x_{i,j}^k$ при выполнении ряда условий.

Условия выбора

$$y_i^k = \begin{cases} 1, \text{ если объект } i \text{ принадлежит кластеру } k, \\ 0 \ \text{ в противном случае, } \ i = \overline{1,n},\ k = \overline{1,m}; \end{cases} \tag{1}$$

$$\sum_{k=1}^{m} y_i^k = 1, \quad i = \overline{1,n}; \tag{2}$$

линеаризующие неравенства

$$0 \le y_i^k + y_j^k - 2x_{i,j}^k \le 1, \quad k = \overline{1,m}, \quad i,\ j = \overline{1,n}, \quad i \ne j,$$

которые при несимметричной матрице расстояний преобразуются в

$$0 \le y_i^k + y_j^k - x_{i,j}^k - x_{j,i}^k \le 1, \quad k = \overline{1,m}, \quad i,j = \overline{1,n}, \quad i \ne j; \tag{3}$$

$$x_{i,j}^k = \begin{cases} 1, \text{ если объекты } i,\ j \text{ принадлежат кластеру } k\colon\ y_i^k = 1,\ y_j^k = 1, \\ 0 \ \text{ в противном случае, } \ i,\ j = \overline{1,n},\ i \ne j,\ k = \overline{1,m}. \end{cases} \tag{4}$$

Добавив в задачу условия, реализующие минимаксный критерий

$$\sum_{j=1}^{n}\sum_{i=1}^{n} c_{i,j}x_{i,j}^{k} \leqslant \lambda, \quad k=\overline{1,m}, \quad i \neq j, \quad \lambda \to \min, \tag{5}$$

имеющий смысл минимизации максимальной по всем кластерам суммы расстояний между всеми объектами каждого кластера, получим искомую формализацию (1)–(5) названной в заглавии раздела задачи.

Кроме критерия (5), в зависимости от содержательного смысла задачи кластеризации, в ряде случаев более приемлемым является аддитивный критерий, который удобно представлять в виде

$$\sum_{j=1}^{n}\sum_{i=1}^{n} c_{i,j}x_{i,j}^{k} = \lambda^{k}, \quad k=\overline{1,m}, \quad i \neq j; \tag{6}$$

$$\sum_{k=1}^{m} \lambda^{k} \to \min. \tag{7}$$

Здесь λ^k — сумма расстояний между всеми парами объектов в кластере k, $k=\overline{1,m}$.

Решение задачи (1)–(4), (6), (7) позволяет находить разбиения множества объектов с заданными расстояниями между всеми парами объектов на заданное число m подмножеств (кластеров), которое гарантирует минимизацию суммы минимальных суммарных расстояний между всеми парами объектов по всем кластерам. Оценим трудоемкость вариантов задачи кластеризации (1)–(5), (1)–(4), (6), (7).

3. Оценки вычислительной сложности задач кластеризации и возможностей релаксации

Формального доказательства NP-трудности представленных постановок задач не приводим, поскольку это опосредует сведение представленных постановок к любой известной задаче с NP-полнотой, что само по себе может оказаться труднорешаемой задачей. Заметим только, что при значительном упрощении ограничивающих условий любой из приведенных выше задач получается NP-трудная задача смешанного целочисленного программирования. В частности, подзадача (1), (2), (5) интерпретируется как NP-трудная задача оптимизации расписаний несвязанных параллельных машин по критерию $C_{\max}$ [15]. Доказательство ее NP-трудности можно найти, например, в работах [15, 16]. С учетом же остальных условий сложность представленных задач увеличивается на много порядков, о чем свидетельствуют вычислительные эксперименты на реальных данных. Покажем, как можно несколько ослабить трудоемкость задач (1)–(5) и (1)–(4), (6), (7).

Для этого используем релаксацию по вспомогательным булевым переменным $x_{i,j}^k$, исключив условия целочисленности (4). Вместо них введем границы изменения непрерывных переменных

$$0 \leqslant x_{i,j}^{k} \leqslant 1, \quad k=\overline{1,m}, \quad i,\, j=\overline{1,n}, \quad i \neq j. \tag{8}$$

Релаксации задач (1)–(5) и (1)–(4), (6), (7) обозначим соответственно как (1)–(3), (5), (8) и (1)–(3), (6)–(8). Отметим снижение числа булевых переменных в релаксированных задачах на величину mn^2. Таким образом, общее число булевых переменных (1)

в задачах (1)–(3), (5), (8) и (1)–(3), (6)–(8) составляет величину mn при наличии mn^2+1 непрерывных переменных (7) и (8) против $m\,n(1+n)$ булевых переменных в задачах (1)–(5) и (1)–(4), (6), (7).

В случае применения представленных задач в практических приложениях разница весьма существенна. Поскольку перспективы разработки приемлемых по точности аппроксимационных эффективных алгоритмов для сформулированных задач весьма туманны, применим для их решения условно экспоненциальные алгоритмы, успешность практического применения которых сильно зависит от фактического числа целочисленных переменных. В этом смысле релаксации (1)–(3), (5), (8) и (1)–(3), (6)–(8) имеют существенные преимущества перед постановками (1)–(5) и (1)–(4), (6), (7).

Докажем эквивалентность постановок (1)–(5) и (1)–(3), (5), (8), а также (1)–(4), (6), (7) и (1)–(3), (6)–(8) для поиска оптимальных значений переменных y_i^k.

Рассмотрим соотношения $x_{i,j}^k = y_i^k y_j^k$ и эквивалентные условия (3) $0 \le y_i^k + y_j^k - x_{i,j}^k - x_{j,i}^k \le 1$, $k = \overline{1,m}$, $i,j = \overline{1,n}$, $i \ne j$, означающие, что $x_{i,j}^k$ истинны только тогда, когда истинны y_i^k и y_j^k. Соответственно, y_i^k и y_j^k истинны одновременно только тогда, когда истинны $x_{i,j}^k$.

Заменим условие целочисленности $x_{i,j}^k$ на условие $0 \le x_{i,j}^k \le 1$ и рассмотрим возможные варианты соотношений (3). Если в оптимальном решении $y_i^k = 1$ и $y_j^k = 1$, то соотношение $x_{i,j}^k = y_i^k\, y_j^k$ выполнится только в случае $x_{i,j}^k = x_{j,i}^k = 1$. Если в оптимальном решении $y_i^k = 1$ и $y_j^k = 0$, то соотношение $x_{i,j}^k = y_i^k\, y_j^k$ выполнится только при $x_{i,j}^k = x_{j,i}^k = 0$. Совершенно аналогичны для случая $y_i^k = 0$ и $y_j^k = 0$, $x_{i,j}^k = x_{j,i}^k = 0$. Таким образом, решения релаксаций (1)–(3), (5), (8) и (1)–(3),(6)–(8) совпадают с решениями задач в исходных постановках (1)–(5) и (1)–(4), (6), (7) соответственно.

Для представленных выше вариантов NP-трудных задач кластеризации не существует теоретически эффективных алгоритмов. Однако для практических применений в достаточной степени разработаны алгоритмы, которые можно именовать как условно экспоненциальные. Это словосочетание означает, что, несмотря на недоказанность их эффективности, данные алгоритмы позволяют за разумное время находить оптимальные либо приближенные к оптимальным решения труднорешаемых задач. В качестве примера может служить алгоритм бинарных отсечений и ветвлений [17, 18]. Его программная реализация была применена для поиска решения сформулированных выше задач кластеризации (1)–(3),(5),(8) и (1)–(3),(6)–(8).

Кроме того, в силу универсальности этих постановок для расчетов вполне применимы стандартные средства смешанного программирования, реализованные в программных системах GUROBI и IBM CPLEX последних версий.

4. Индексирование документов, описывающих прецеденты принятия решений в области ИТ-консультирования

В качестве примера применения предлагаемого подхода рассмотрим задачу кластеризации документов, являющихся прецедентами принятия решений в области консультирования по вопросам ИТ-поддержки. В работе [19] построена онтология в области ИТ-поддержки, иерархическая основа которой может быть использована в качестве таксономии концептов для индексирования прецедентов. На рисунке представлен фрагмент таксономии (с тремя связями соответственно с весами $\nu_1 = 0.6$, $\nu_2 = 0.1$ и $\nu_3 = 0.3$ от документа Precedent к концептам, являющимся транзитивным замыкани-

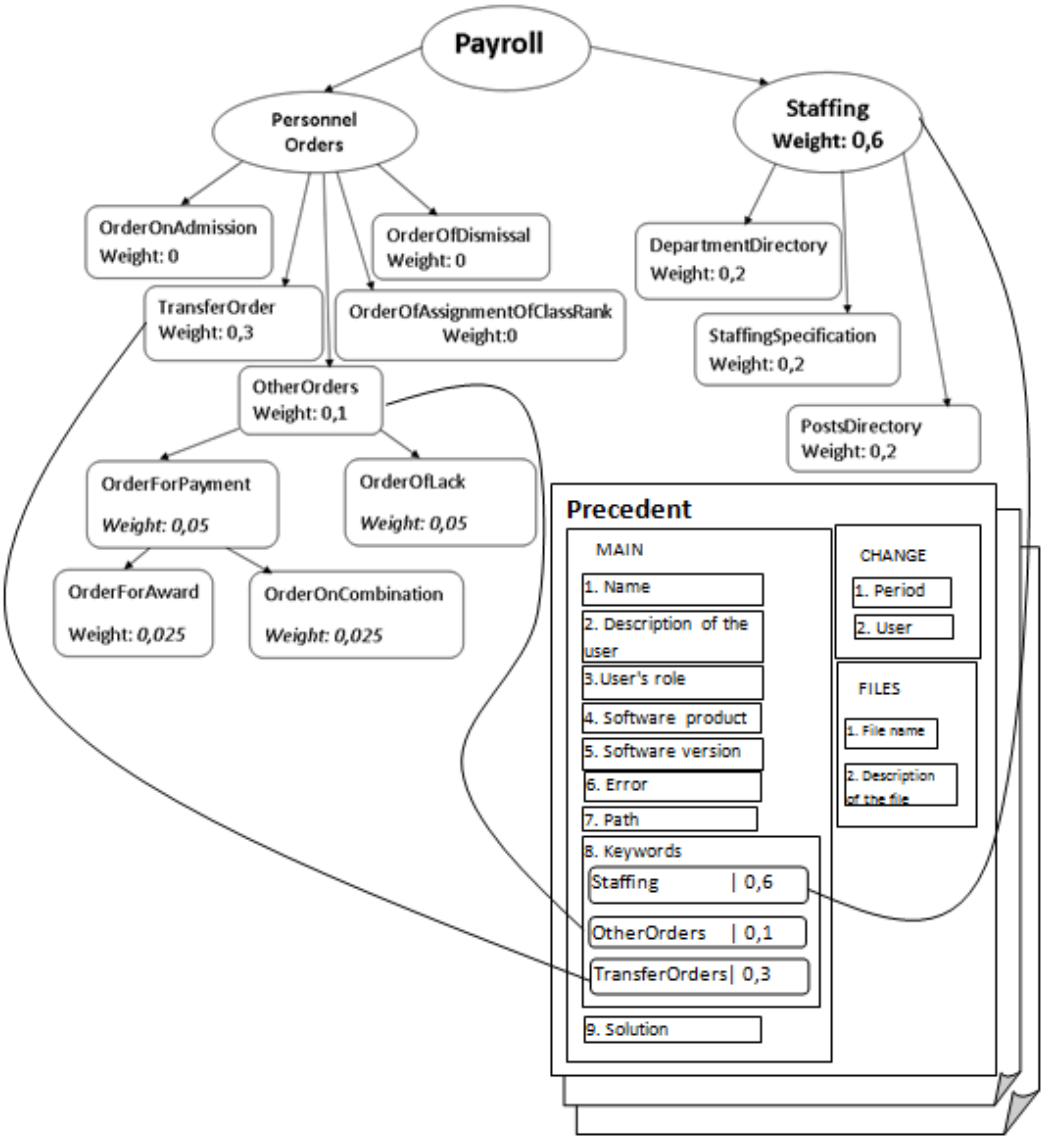

Рис. 1. Формирование семантической матрицы связи документа Precedent с таксономией
Fig. 1. Generating the semantic matrix of the relationships between the Precedent document and the taxonomy

ем концепта Payroll. Для данного фрагмента таксономии имеем следующий подвектор весов $\widetilde{\mathbf{w}} = \begin{pmatrix} 0 & 0.3 & 0.025 & 0.025 & 0.05 & 0 & 0 & 0.2 & 0.2 & 0.2 \end{pmatrix}^T$.

Таким образом индексируются все документы с соответствующими концептами онтологии [19] (120 документов, 20 терминальных концептов онтологии, общих для рассматриваемого множества документов). В результате получена матрица семантических связей 120×20 со значениями весов w_j, $j = \overline{1, J}$, в качестве строк, которая далее использовалась в качестве матрицы данных для проведения кластерного анализа.

В работе [20] данная матрица применена для решения задачи выбора наиболее близких концептов таксономии методом комбинированного использования корреляционных мер для структурирования пространства концептов и перепроектирования онтологии. В настоящей работе мы используем полученную семантическую матрицу для решения задачи кластеризации прецедентов.

5. Результаты применения метода оптимальной кластеризации

Как следует из описанных выше условий, задача m-кластеризации содержит 120 объектов, 20 показателей (терминальных концептов онтологии) при $m = 10$. Оценим размерность реализации задачи (1)–(3), (6)–(8). Число булевых переменных y_i^k составляет $mn = 1200$, непрерывных переменных $x_{i,j}^k$ и λ^k: $m\,n^2+m = 144\,010$ при соответствующем выражениям (2), (3), (6), (8) количестве ограничений. Немалая размерность реализации NP-трудной задачи смешанного программирования предопределила вычислительные сложности и потребовала при решении посредством оптимизатора IBM CPLEX применения ограничений на время счета и использования декомпозиции задачи.

Прямое применение 10-кластеризации с ограничением времени счета до приемлемого значения привело к следующему результату. Вектор суммарных расстояний между объектами внутри кластеров: $\boldsymbol{\lambda} = (45.262; 63.264; 49.271; 51.967; 48.968; 45.81; 54.579; 25.674; 49.976; 61.137)$ с общим значением критерия — суммы евклидовых расстояний между всеми парами объектов $\sum \lambda^k = 495.908$. Количество элементов в кластерах (12; 13; 12; 12; 12; 11; 12; 8; 11; 17). Данный результат не гарантировал какой-либо меры близости к оптимуму, что опосредовало необходимость снижения размерности и применения декомпозиции. Декомпозиция проведена по двухэтапной иерархической схеме:

1) выделение двух кластеров (непересекающихся подмножеств объектов);

2) последовательная 5-кластеризация каждого из сформированных на первом этапе подмножеств.

Конечный результат кластеризации с декомпозицией по двухэтапной схеме оказался примерно на 20 % лучше исходного: $\boldsymbol{\lambda} = (30.94; 44.18; 44.37; 32.76; 42.65; 40.88; 39.92; 46.12; 45.97; 32.90)$ с общим значением критерия $\sum \lambda^k = 400.68$ и числом элементов в кластерах (12; 13; 13; 10; 11; 11; 11; 16; 12; 11). Такой результат объясняется возможностью существенно большего приближения к локальным для подзадач оптимумам при снижении размерности.

Таким образом, результаты расчетов подтвердили перспективность оптимизационного подхода для кластеризации множеств объектов различной природы и выявили проблемы вычислительного характера. В частности, стала очевидна необходимость развития специальных вычислительных инструментов (алгоритмов оптимизации, опирающихся на специфику сформулированных задач) и совершенствования самих формальных постановок для снижения общей трудоемкости вычислений.

Благодарности. Работа выполнена при финансовой поддержке Министерства науки и высшего образования в рамках Госзадания (проект № FSUN-2020-0009).

Список литературы

[1] **Bolelli L., Ertekin S., Lee G.C.** Clustering scientific literature using sparse citation graph analysis. Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). Germany; 2006: 30–41.

[2] **Sedding J., Kazakov D.** WordNet-based text document clustering. Proc. COLLING-2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data. 2004: 104–113.

[3] **Sebastiani F.** Machine learning in automated text categorization. ACM Computing Surveys. 2002; (34):1–47.

[4] **Boubekeur F., Azzoug W.** Concept-based indexing in text information retrieval. Intern. J. of Computer Science and Inform. Technology (IJCSIT). 2013; 5(1):119–136.

[5] **Scott S., Stan M.** Text classification using wordnet hypernyms. WordNet@ACL/COLING, 1998: 45–51.

[6] **Rajman M., Andrews P., Almenta M.D., Seydoux F.** Conceptual document indexing using a large scale semantic dictionary providing a concept hierarchy. Proc. Applied Stochastic Models and Data Analysis (ASMDA 2005), France. 2005: 88–05.

[7] **Лукашевич Н.В.** Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа: автореферат дис. ... доктора технических наук. М.: Всерос. ин-т науч. и техн. информ. (ВИНИТИ) РАН; 2014: 33.

[8] **Barresi S., Nefti-Meziani S., Rezgui Y.** A concept based indexing approach for document clustering. 2008 IEEE Intern. Conf. on Semantic Computing, Santa Clara, CA. 2008: 26–33. DOI:10.1109/ICSC.2008.75.

[9] **Kelmanov A.V., Pyatkin A.V.** NP-Difficulty of some Euclidean problems of partitioning a finite set of points. J. Computational Mathematics and Mathematical Physics. 2018; 58(5): 852–856.

[10] **Кельманов А.В., Пяткин А.В.** О сложности некоторых задач кластерного анализа векторных последовательностей. Дискретный анализ и исследование операций. 2013; 20(2):47–57.

[11] **Мезенцев Ю.А., Разумникова О.М., Тарасова И.В., Трубникова О.А.** О некоторых задачах кластеризации больших данных по минимаксным и аддитивным критериям, применение в медицине и нейрофизиологии. Информационные технологии. 2019; 25(10):602–608.

[12] **MacQueen J.** Some methods for classification and analysis of multivariate observations. Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967; (1): 281–297.

[13] **Кохонен Т.** Самоорганизующиеся карты. М.: БИНОМ. Лаборатория знаний; 2008: 655.

[14] **Shen F., Ogurab T., Hasegawa O.** An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks. 2007; 20(8):893–903.

[15] **Pinedo M.** Scheduling. Theory, Algorithms, and Systems. Third Edition. New York: Springer; 2008: 672. DOI:10.1007/978-0-387-78935-4.

[16] **Lenstra J.K., Shmoys D.B., Tardos E.** Approximation algorithms for scheduling unrelated parallel machines. Mathematical Programming. 1987; (46):259–271.

[17] **Mezentsev Yu.A.** Binary cut-and-branch method for solving linear programming problems with boolean variables. Proc. 9th Intern. Conf. on Discrete Optimization and Operations Research and Scientific School. 2016; (1623):72–85.

[18] **Mezentsev Yu.** Binary cut-and-branch method for solving mixed integer programming problems. Constructive Nonsmooth Analysis and Related Topics (Dedicated to the Memory of V.F. Demyanov), CNSA, 2017. `https://doi.org/10.1109/cnsa.2017.7973989`

[19] **Avdeenko T.V., Makarova E.S.** The case-based decision support system in the field of IT-consulting. Journal of Physics: Conference Series. 2017; (803): 012008.

[20] **Timofeeva A.Y., Avdeenko T.V., Makarova E.S., Murtazina M.S.** Combined use of correlation measures for selecting semantically close concepts of the ontology. CEUR Workshop Proceedings. 2018; (2212):349–358.

Вычислительные технологии, 2020, том 25, № 3, с. 99–110. © ИВТ СО РАН, 2020 ISSN 1560-7534
Computational Technologies, 2020, vol. 25, no. 3, pp. 99–110. © ICT SB RAS, 2020 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2020.25.3.011

Document clustering based on the semantic matrix of relationships for conceptual indexing

Avdeenko Tatiana V.*, Mezentsev Yuriy A.
Novosibirsk State Technical University, 630073, Novosibirsk, Russia
*Corresponding author: Avdeenko Tatiana V., e-mail: avdeenko@corp.nstu.ru
Received March 10, 2020, revised March 20, 2020, accepted April 14, 2020

Abstract

Purpose. The purpose of this work is to develop a method for document clustering based on conceptual indexing with the help of knowledge taxonomy.

Methodology. Solving the problem of document clustering involves two fundamental stages. The first stage is preprocessing of a text document and representing it as a data table suitable for subsequent application of data analysis methods. The second stage is actually the optimization of clustering algorithm, which allows achieving optimal partitioning of the document collection in order to achieve, on the one hand, compactness of clusters, on the other hand, distinctness of clusters. We suggest a new approach to conceptual indexing of documents by transformation of a set of key terms to a weighted set of concepts for a certain hierarchical knowledge model of the application domain. The semantic matrix of document relationships with taxonomy concepts obtained as a result of the approach can be used as a data matrix for solving the clustering problem. For this purpose, we propose an original approach that uses the formalization of an NP-hard mixed programming problem, decomposition, and step-by-step solution that reduces its complexity.

Results. As an example of applying the proposed approach, we consider the problem of clustering of 120 documents using 20 features that are terminal concepts of taxonomy. The result of direct clustering across 10 clusters did not guarantee closeness to the optimum, which caused the need for decomposition. The decomposition was carried out according to a two-stage hierarchical scheme: the first stage is the allocation of 2 clusters, the second stage is a sequential 5-clustering of each of the subsets formed at the first stage. The final result of clustering with two-step decomposition was about 20 percent better than the original one.

Findings. The results of calculations confirmed the prospects of an optimization approach for clustering documents using a semantic matrix of relationships and revealed computational problems. In particular, the necessity of developing special computing tools and improving the formal statements themselves for reducing the overall complexity of calculations was revealed.

Keywords: document clustering, conceptual indexing, taxonomy, ontology, mixed integer programming, NP-hard problem.

Citation: Avdeenko T.V., Mezentsev Yu.A. Document clustering based on the semantic matrix of relationships for conceptual indexing. Computational Technologies. 2020; 25(3):99–110. (In Russ.)

Acknowledgements. The research is supported by Ministry of Science and Higher Education of RF (project No. FSUN-2020-0009).

References

1. Bolelli L., Ertekin S., Lee G.C. Clustering scientific literature using sparse citation graph analysis. Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). Germany; 2006: 30–41.

2. Sedding J., Kazakov D. WordNet-based text document clustering. Proc. COLLING-2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data. 2004: 104–113.
3. Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys. 2002; (34):1–47.
4. Boubekeur F., Azzoug W. Concept-based indexing in text information retrieval. Intern. J. of Computer Science and Information Technology (IJCSIT). 2013; 5(1):119–136.
5. Scott S., Stan M. Text classification using wordnet hypernyms. WordNet@ACL/COLING, 1998: 45–51.
6. Rajman M., Andrews P., Almenta M.D., Seydoux F. Conceptual document indexing using a large scale semantic dictionary providing a concept hierarchy. Proc. Applied Stochastic Models and Data Analysis (ASMDA 2005), France. 2005: 88–05.
7. Lukashevich N.V. Modeli i metody avtomaticheskoy obrabotki nestrukturirovannoy informatsii na osnove bazy znaniy ontologicheskogo tipa [Models and methods for automatic processing of unstructured information based on an ontological knowledge base]. Moscow: VINITI RAN; 2014: 33. (In Russ.)
8. Barresi S., Nefti-Meziani S., Rezgui Y. A concept based indexing approach for document clustering. 2008 IEEE Intern. Conf. on Semantic Computing, Santa Clara, CA. 2008: 26–33. DOI:10.1109/ICSC.2008.75.
9. Kelmanov A.V., Pyatkin A.V. NP-Difficulty of some Euclidean problems of partitioning a finite set of points. J. Computational Mathematics and Mathematical Physics. 2018; 58(5):852–856.
10. Kelmanov A.V., Pyatkin A.V. On the complexity of some vector sequence clustering problems. J. of Applied and Industrial Mathematics. 2013; 7(3):363–369.
11. Mezentsev Yu.A., Razumnikova O.M., Tarasova I.V., Trubnikova O.A. On some problems of big data clustering by minimax and additive criteria, application in medicine and neurophysiology. Information Technologies. 2019; 25(10):602–608. (In Russ.)
12. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967; (1):281–297.
13. Kohonen T. Self-organizing maps (Third extended edition). New York; 2001: 501.
14. Shen F., Ogurab T., Hasegawa O. An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks. 2007; 20(8):893–903.
15. Pinedo M. Scheduling. Theory, Algorithms, and Systems. Third Edition. New York: Springer; 2008: 672. DOI:10.1007/978-0-387-78935-4.
16. Lenstra J.K., Shmoys D.B., Tardos E. Approximation algorithms for scheduling unrelated parallel machines. Mathematical Programming. 1987; (46):259–271.
17. Mezentsev Yu.A. Binary cut-and-branch method for solving linear programming problems with boolean variables. Proc. 9th Intern. Conf. on Discrete Optimization and Operations Research and Scientific School. 2016; (1623):72–85.
18. Mezentsev Yu. Binary cut-and-branch method for solving mixed integer programming problems. Constructive Nonsmooth Analysis and Related Topics (Dedicated to the Memory of V.F. Demyanov), CNSA, 2017. https://doi.org/10.1109/cnsa.2017.7973989
19. Avdeenko T.V., Makarova E.S. The case-based decision support system in the field of IT-consulting. Journal of Physics: Conference Series. 2017; (803): 012008.
20. Timofeeva A.Y., Avdeenko T.V., Makarova E.S., Murtazina M.S. Combined use of correlation measures for selecting semantically close concepts of the ontology. CEUR Workshop Proceedings. 2018; (2212):349–358.