Системы управления и информационные технологии. 2015. № 3-1 (61). С. 134-139.

УДК 004.822

Выделение семантического ядра на основе матрицы корреспонденций термов

Д.В. Бондарчук, Г.А. Тимофеева
Уральский государственный университет путей сообщения

Аннотация
Выделение семантического ядра широко используется для компактного представления документов при решении задач классификации и интеллектуального поиска. Для выделения семантического ядра вводится матрица корреспонденций термов, отражающая взаимосвязи между документами, проводится её сингулярное разложение. Результаты сравниваются с сингулярным разложением терм-документной матрицы, которое используется в латентно-семантическом анализе.
Ключевые слова: интеллектуальный поиск, семантическое ядро, матрица корреспонденций термов.

Isolation of the semantic kernel based on the matrix of terms correspondence

D.V. Bodarchuk, G.A. Timofeeva
Ural State University of Railway Transport

The isolation of the semantic core is widely used for the compact presentation of the documents in solving problems of classification and intelligent search. The matrix of terms correspondence reflecting the relationship between documents is introduced. Its singular value decomposition is carried out for solving the problem of the semantic core isolation. The results are compared with the singular value decomposition of a term-document-matrix, which is used in latent semantic analysis.
Keywords: intelligent search, semantic core, matrix of terms correspondence.

Введение

Векторная модель представления документа является упрощенной моделью представления документа, которая учитывает частоту встречаемости слов, но не учитывает их грамматику, семантику и прочие особенности текстов на естественном языке. С лингвистической точки зрения данное допущение является критическим, поскольку в большинстве случаев смысл документа напрямую зависит от этих особенностей текста.

Сбор семантического ядра на основе векторной модели - это относительно новый подход решения проблем обработки текстов на естественном языке [1]. Одним из ключевых достоинств данного подхода является его модульность: разграничение собственно алгоритма анализа текстовых данных от статистического анализа частоты встречаемости терминов, необходимого на предварительном этапе. Кроме того, сами ядра имеют модульную структуру, что позволяет с помощью простых правил строить более сложные семантические ядра из более простых таким образом, чтобы они не выходили за границы семантического пространства. Основная идея применения методов, основанных на сборе семантического ядра - переход к новому семантическому пространству, размерность которого меньше размерности исходного пространства и проведение интеллектуального анализа данных в котором легче. Такой подход применяется в частности при решении задачи подбора вакансий [2,3].

Постановка задачи

Рассматривается задача представления информации с целью ее можно использования в качестве обучающей базы для текстового классификатора, построенного на основе векторной модели представления знаний. На этом этапе взаимосвязь слов внутри документа, а также проблемы синонимии и полисемии не рассматриваются.

В большинство современных подходов взаимосвязь между термами рассчитывается только с помощью оценки их распределения по всему набору документов, что свою очередь может привести к некоторому снижению качества обучения.

Семантическое ядро - это подборка понятий, имеющих существенное значение для данной предметной области. Точное определение семантического ядра зависит от области применения. Так, в лингвистике, семантическим ядром называют "не упрощаемое замкнутое подмножество языка", подразумевая при этом скорее смысловую составляющую языка, а не грамматические конструкции.

Прежде, чем переходить к статистическому анализу текстовых данных, необходимо произвести ряд действий для упрощения статистической обработки. *Стемминг* - это процесс нахождения основы слова для заданного исходного слова при этом основа слова не обязательно совпадает с морфологическим корнем слова. Алгоритмы стемматизации (стеммеры) применяются в поисковых системах для обобщения поискового запроса пользователя. Наиболее удачный алгоритм стемминга — стеммер Портера, оригинальная версия которого предназначена для английского языка [4]. Алгоритм не использует баз основ слов, а, применяя последовательно ряд правил, отсекает части слов, основываясь на особенностях языка, в связи с чем работает быстро, но не всегда

безошибочно.

Термом будем называть слово, обработанное с помощью стеммера Портера, и не содержащееся в списке стоп-слов. Под *стоп-словами,* понимаются слова, содержащиеся почти в каждом тексте и не несущие никакой смысловой нагрузки («кто», «куда», «ли», «лучше», «между» и т.д.) Отметим, что для каждой предметной области список стоп-слов может быть свой.

Одним из наиболее эффективных методов классификации текстов является метод латентно-семантического анализа (ЛСА)[5, 6, 7]. Он позволяет выявлять значения слов с учетом контекста их использования путем обработки большого объема текстов. Модель представления текста, используемая в ЛСА, во многом схожа с восприятием текста человеком.

В данной статье предлагается другой метод выделения семантического ядра, основанный построении и анализе матрицы корреспонденций термов (МКТ), которая отражает взаимосвязи между термами.

Матрица корреспонденций термов

Для решения проблемы выделения семантического ядра введем новое понятие - *матрица корреспонденций термов*.

Определение. Матрица корреспонденций термов $G=\{g_{ij}\}$ - это квадратная матрица, элементами которой являются коэффициенты $g_{ij,}$ отражающие близость i-го и j-го термов, для которых выполняются следующие условия:

1. $g_{ij,} = g_{ji,}$
2. $-1 \le g_{ij} \le 1$ для всех i и j.
3. $g_{ij,} = 0$ при отсутствии взаимосвязи между термами.

В качестве меры близости можно рассматривать:

- скалярные произведения векторов, соответствующих нормированным векторам термов;
- корреляцию между векторами-термами;
- меры Дайса (*Dice measure*) и Джакарта (*Jaccard measure*) [8].

Мера близости термов может корректироваться с учетом их семантической близости.

Основное назначение матрицы G – отображение взаимосвязей термов внутри документов, построенное на основе знаний частоте об их совместных употреблениях. На рисунке 1 изображен случай, когда термы t_1 и t_2 совместно встречаются в документе d_2, а термы t_2 и t_3 - в документе d_1. Таким образом, термы t_1 и t_3 так же связаны между собой через терм t_2.

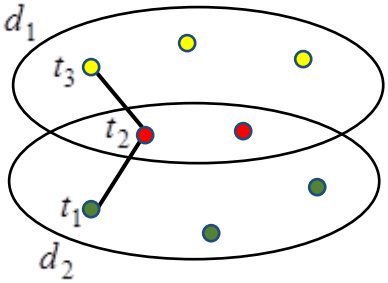

Рисунок 1. Иллюстрация взаимосвязей термов

Латентно-семантический анализ

ЛСА отображает документы и отдельные слова в так называемое «семантическое пространство», в котором и производятся все дальнейшие сравнения. Для построения семантического пространства используется терм-документная матрица, отражающая количество появлений терминов (термов) в документах.

При этом делаются следующие предположения:

1) Документ это просто набор слов. Порядок слов в документах игнорируется. Важно только то, сколько раз то или иное слово встречается в документе.

2) Семантическое значение документа определяется набором слов, которые, как правило, идут вместе.

3) Каждое слово имеет единственное значение. Это, безусловно, сильное упрощение, но именно оно делает проблему разрешимой.

В качестве исходной информации используется терм-документная матрица X, которая описывает частоту термов.

Предположим у нас есть некоторая обучающая выборка текстов. Представим её в виде матрицы X, строками которой являются x_i - вектора термов, n - количество термов. Вектор терма t_i представляет собой вектор-строку:

$$x_i = \{tf(t_i, d_1), tf(t_i, d_2), \ldots, tf(t_i, d_m)\}, \quad (1)$$

где d_j – j-ый документ из обучающей выборки, $tf(ti,dj)$ - частота встречаемости терма t_i в документе d_j (*term frequency*), m - количество документов, содержащихся в обучающей выборке. Частота встречаемости терма в документе равна числу вхождений терма t_i в документ d_j:

$$x_{ij} = tf(t_i, d_j), \quad (2)$$

Матрица X называется *терм-документной матрицей*.

При использовании классической векторной модели, предложенной Салтоном [1], представление каждого документа содержит в себе лишь статистическую информацию о появлении термов в коллекции. Стандартный ЛСА не предусматривает никакой предварительной работы с исходной матрицей, однако ее преобразование может значительно повысить эффективность данного метода. Предлагается произвести с матрицей следующие действия, которые позволят существенно уменьшить ее размерность [7]:

- удалить строки, соответствующие стоп-словам;
- удалить строки, соответствующие редким словам, не встречающимся ни в одном тексте из выборки более одного раза;
- привести все словоформы к исходной форме, например с помощью операции стемминга (процесс нахождения основы слова);
- из текстов некоторых тематик полезно удалить имена собственные, которые, так же как и стоп-символы не несут в себе никакой смысловой нагрузки;
- из текстов некоторых тематик имеет смысл удалить всю цифровую информацию (числительные, цифры).

В результате применения перечисленных действий количество строк в исходной матрице уменьшается в среднем на 60-70%, при больших выборках этот факт значительно увеличивает скорость обучения классификатора. Отметим, что, если целью ЛСА является информационный поиск, то удаление шумовой информации может повлечь за собой снижение результативности поиска. После того как вся возможная шумовая информация удалена можно приступать к следующему шагу ЛСА - сингулярному разложению терм-документной матрицы.

Сингулярное разложение матриц

Сингулярное разложение - это математическая операция, представляющая матрицу A размера $m \times n$ в виде произведения

$$A = USV^T \tag{3}$$

где U и V - ортогональные матрицы размера $m \times m$ и $n \times n$ соответственно, т.е. $UU^T = E$, $VV^T = E$, S - прямоугольная диагональная матрица размера $m \times n$. Под прямоугольной диагональной матрицей понимается матрица $S = \{s_{ij}\}$ такая, что $s_{ij} = \lambda_i$ при $j = i \leq \min\{m, n\}$ и $s_{ij} = 0$ в остальных случаях.

Как известно [9], для любой вещественной $m \times n$ матрицы A существует сингулярное разложение (3). Более того, матрицы U и V можно

подобрать таким образом, чтобы диагональные элементы S были расположены по убыванию:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0 \quad (4)$$

где r - ранг матрицы A, то есть сингулярный коэффициент в строке матрицы S всегда больше, либо равен коэффициенту в строке ниже.

В частности, в случае, если матрица A квадратная и невырожденная, то:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0. \quad (5)$$

Общий алгоритм сингулярного разложения можно представить следующей последовательностью шагов:

1. Вычисление матрицы AA^T, нахождение собственных чисел и собственных векторов матрицы AA^T.

2. Вычисление матрицы $A^T A$, нахождение собственных векторов матрицы $A^T A$.

3. Составление матриц U, V и S.

При проведении ЛСА в сингулярном разложении терм-документной матрицы X, определяемой соотношениями (1)-(2), обычно оставляют только ненулевые строки и столбцы матрицы сингулярных коэффициентов S, при этом отбрасывают соответствующие строки матрицы U и столбцы матрицы V^T. Таким образом, получают разложение

$$X = USV^T, \quad (6)$$

где U - $n \times r$ матрица, S - $r \times r$ матрица, V^T - $r \times m$ матриц, где r - ранг матрицы S. На рисунке 2 изображено схематично сингулярное разложение терм-документной матрицы.

Матрица левых сингулярных векторов U в методе латентно-семантического анализа называется матрицей размерностей термов (англ. *term by dimension matrix*); значения λ_i на диагонали матрицы S называются *сингулярными коэффициентами* матрицы X. Сингулярные коэффициенты всегда неотрицательны. Матрица правых сингулярных векторов V в методе латентно-семантического анализа V называется матрицей размерностей документов (англ. *document by dimension matrix*).

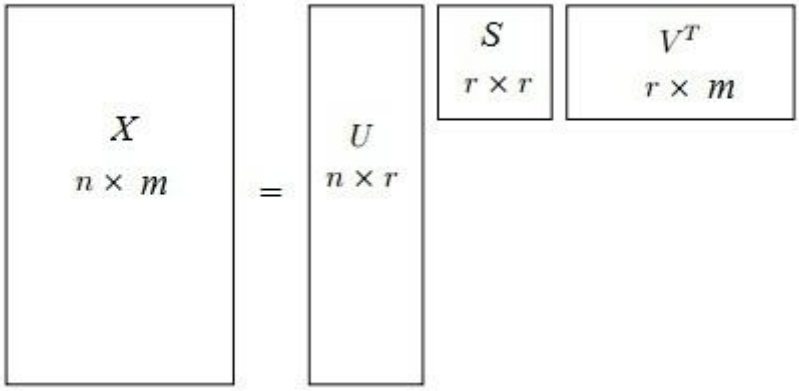

Рисунок 2. Размерности матриц в ЛСА

Основная идея ЛСА состоит в том, что матрица X_k, содержащая только k первых линейно независимых компонент разложения (6) терм-документной матрицы X, отражает основную структуру различных зависимостей, присутствующих в исходной матрице [6]. Матрица S_k находится оставлением в диагональной матрице S только первых k строк с наибольшими элементами, при этом разложение полученной матрицы имеет вид:

$$X_k = U_k S_k V_k^T \tag{7}$$

Выбор k зависит от поставленной задачи и подбирается эмпирически. Если выбранное значение k слишком велико, то метод теряет свою мощность и приближается по характеристикам к стандартным векторным методам. Слишком маленькое значение k не позволяет улавливать различия между похожими термами или документами. Если же необходимо выбирать значение k автоматически, то можно, например, установить пороговое значение сингулярных коэффициентов λ_k и отбрасывать все строки и столбцы, соответствующие сингулярным коэффициентам, не превышающим данное пороговое значение. При этом происходит так называемое *семантическое сглаживание*, которое позволяет избежать изолированности термов при вычислении сходства между документами.

Дополнительную сложность для любого классификатора текстов представляют синонимия и полисемия. Синонимия - одинаковость или сходство значения различных слов или других однородных языковых единиц. Полисемия - многозначность, многовариантность, то есть наличие у слова (единицы языка, термина) двух и более значений, исторически обусловленных или взаимосвязанных по смыслу и происхождению. Такие слова еще называют омонимами.

Сингулярное разложение матрицы корреспонденций термов

Как было указано выше, в качестве матрицы корреспонденций термов можно брать различные матрицы. Рассмотрим подробнее случай, когда в качестве МКТ взята матрица, полученная из произведений нормированных векторов-термов.

Построим нормированную терм-документную матрицу $Y=\{y_{ij}\}$, где

$$y_{ij} = \frac{x_{ij}}{\sum_i x_{ij}} = \frac{x_{ij}}{n_j} \tag{8}$$

здесь x_{ij} - число вхождений слова в документ, а n_j - общее количество слов в документе d_j. Через y_i обозначим вектор-строку

$$y_i = \{y_{i1},\ldots,y_{im}\} \ . \tag{9}$$

Построим матрицу, состоящую из всех возможных скалярных произведений векторов термов y_i, определяемых по формулам (8) - (9):

$$G = ((y_i, y_j))_{i,j=1}^{n} = YY^T . \tag{10}$$

Матрица, составленная из скалярных произведений, называется матрицей Грама. Очевидно, что эта матрица является симметричной и неотрицательно определенной.

Применим сингулярное разложение к матрице корреспонденций термов G, определенной по формуле (10). Сингулярное разложение аналогично процедуре из метода латентно-семантического анализа, однако, в данном случае вместо терм-документной матрицы X, которая используется в ЛСА, будем использовать сингулярное разложение матрицы корреспонденций термов G, которая отражает взаимосвязь термов в корпусе.

Матрица корреспонденций термов (МКТ) связана с моделью представления знаний через терм-документную матрицу, однако ее назначение отлично. Разложение корреспонденций термов может быть получено из разложения нормированной терм-документной матрицы Y следующим образом.

Утверждение 1. Сингулярное разложение матрицы корреспонденций термов G, определенной по формуле (10) *имеет вид*:

$$G = TZT^T \tag{11}$$

где T – ортогональная матрица левых сингулярных векторов в разложении нормированной терм-документной матрицы Y, матрица Z – диагональная матрица размера, на диагонали которой стоят $(\lambda_i)^2$, λ_i - *сингулярные коэффициенты разложения матрицы Y.*

Утверждение легко проверяется. Матрицу корреспонденций термов *G* можно представить следующим образом:

$$G = YY^T = TS_Y D^T D S_Y{}^T T^T .$$

С учетом ортогональности матрицы *D* получаем:

$$G = TS_Y S_Y{}^T T^T = TZT^T .$$

Часть матрицы, содержащая только *k* линейно независимых компонент, будет отражать основную структуру зависимостей, присутствующих в исходной матрице. Таким образом, термы, которые чаще встречаются в корпусе, получат в результате разложения более высокие сингулярные коэффициенты. Усеченную матрицу *G* до размерности *k* обозначим G_k:

$$G_k = T_k Z_k T_k^T \qquad (12)$$

Замечание. Применение сингулярного разложения к стандартной терм-документной матрице *X* и к нормированной терм-документной матрице *Y* приводит, вообще говоря, к различным семантическим пространствам, поэтому преобразование T_k полученное с использованием МКТ не совпадает с матрицей U_k стандартного ЛСА.

Пример 1. Пусть терм-документная матрица *X* равна:

$$X = \begin{pmatrix} 100 & 100 & 100 & 0 & 0 \\ 1 & 3 & 1 & 0 & 1 \\ 2 & 0 & 3 & 1 & 0 \\ 4 & 1 & 1 & 1 & 1 \\ 4 & 3 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Сингулярное разложение терм-документной матрицы имеет вид (7), где:

$$U = \begin{pmatrix} -1.00 & -0.03 & 0.01 & 0.01 & 0.01 \\ -0.02 & 0.15 & 0.40 & -0.76 & -0.48 \\ -0.02 & -0.19 & -0.66 & -0.12 & -0.43 \\ -0.02 & 0.58 & -0.48 & 0.05 & -0.30 \\ -0.02 & 0.77 & 0.18 & 0.17 & 0.12 \\ -0.01 & 0.11 & -0.37 & -0.61 & 0.69 \end{pmatrix},$$

$$S = \begin{pmatrix} 173.3 & 0 & 0 & 0 & 0 \\ 0 & 3.949 & 0 & 0 & 0 \\ 0 & 0 & 3.411 & 0 & 0 \\ 0 & 0 & 0 & 1.557 & 0 \\ 0 & 0 & 0 & 0 & 0.048 \end{pmatrix}.$$

Запишем нормализованную матрицу

$$Y = \begin{pmatrix} 0.333 & 0.333 & 0.333 & 0 & 0 \\ 0.166 & 0.5 & 0.166 & 0 & 0.166 \\ 0.333 & 0 & 0.5 & 0.166 & 0 \\ 0.5 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.5 & 0.375 & 0 & 0 & 0.125 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \end{pmatrix},$$

Матрицы T и S_Y ее сингулярного разложения имеют вид:

$$T = \begin{pmatrix} -0.45 & -0.15 & -0.46 & 0.31 & -0.56 \\ -0.38 & -0.52 & -0.35 & -0.53 & 0.43 \\ -0.38 & 0.53 & -0.41 & 0.31 & 0.35 \\ -0.44 & 0.09 & 0.52 & 0.25 & 0.43 \\ -0.41 & -0.41 & 0.44 & 0.20 & -0.25 \\ -0.38 & 0.50 & 0.21 & -0.66 & -0.36 \end{pmatrix},$$

$$S_Y = \begin{pmatrix} 1.214 & 0 & 0 & 0 & 0 \\ 0 & 0.573 & 0 & 0 & 0 \\ 0 & 0 & 0.38 & 0 & 0 \\ 0 & 0 & 0 & 0.274 & 0 \\ 0 & 0 & 0 & 0 & 0.007 \end{pmatrix}.$$

В приведенном примере результаты разложения нормированной и ненормированной матриц – существенно разные, так как им соответствуют разные матрицы линейных преобразований U и T. Так же различными будут размерности усеченных матриц, поскольку будет отброшено разное количество термов. В первом случае (без нормировки) будут отброшены все, кроме первого терма, во втором (с нормировкой) - все, кроме первого и второго. Это в свою очередь приведет к образованию разных семантических пространств и получению разных результатов обучения.

Такой результат в примере получен, прежде всего, за счет существенного различия в количествах термов в документах. Нетрудно проверить, что при одинаковом числе термов в документах оба разложения дадут один и тот же результат.

Таким образом, матрица T_k является матрицей линейного преобразования, переводящей вектора из исходного пространства в семантическое. Вычислительные эксперименты показали, что при проведении интеллектуального анализа данных с использованием матрицы корреспонденций термов внутренние зависимости отражаются более явно, что приводит к большему снижению размерности семантического пространства. Отметим, что предлагаемый метод выделения семантического ядра на основе сингулярного разложения МКТ аналогичен методу главных компонент в статистическом анализе, если в качестве меры близости термов взять коэффициент корреляции.

Использование матрицы корреспонденций термов для выделения семантического ядра является более гибким методом по сравнению с латентно-семантическим анализом, поскольку позволяет использовать различные меры близости термов. Использование нестандартных мер может быть необходимо, например, при обработке большого количества текстов из разных предметных областей в рамках обучения одного классификатора.

Выводы

В результате применения метода отбора семантического ядра с помощью матрицы корреспонденций термов происходит переход в новое семантическое пространство, размер которого оказывается значительно меньше исходного, что в свою очередь приводит к увеличению общей производительности текстового классификатора. Кроме прочего, малозначимые термы пропадают из общего списка термов, что приводит к уменьшению количества ложных выводов классификатора.

Основным недостатком предлагаемого метода является значительная сложность вычислений. Используемая матрица корреспонденций термов, как правило, несколько больше, чем матрица терм-документальная, использующаяся в ЛСА. Так МКТ сильно разрежена, то предварительная обработка обучающей выборки: удаление стоп-слов, стеммизация значительно улучшает работу алгоритма.

Список литературы

1. Salton G., Wong A., Yang C.-S. A vector space model for automatic indexing// Communications of the ACM, 1975, 18 (11). P. 613-620.
2. Бондарчук Д.В. Интеллектуальный метод подбора персональных

рекомендаций, гарантирующий получение непустого результата. // Информационные технологии моделирования и управления. 2015. №2(92). С. 130-138.
3. Бондарчук Д.В. Выбор оптимального метода интеллектуального анализа данных для подбора вакансий // Информационные технологии моделирования и управления. 2013. №6(84). С. 504-513.
4. Porter M. An algorithm for Suffix Stripping // Program, 1980, 14(3). P. 130-137.
5. Deerwester S., Dumais S.T., Furnas G.W., Landauer T. K., Harshman R. Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science, 1999, 41 (6). P. 391-407.
6. Landauer T., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis // Discourse Processes, 1998, 25: P. 259–284.
7. Бондарчук Д.В. Использование латентно-семантического анализа в задачах классификации текстов по эмоциональной окраске // Бюллетень результатов научных исследований. 2012. №2 (3). С. 146-151.
8. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
9. Тыртышников Е.Е. Матричный анализ и линейная алгебра. - М.: Физматлит, 2007. 480 с.