Алгоритмы интеллектуального анализа данных на основе категориальных векторов

Д.В. Бондарчук
Руководитель: проф., д.ф.-м.н. Г.А. Тимофеева

Основные алгоритмы интеллектуального анализа данных

- Метод латентно-семантического анализа (Landauer T., Deerwester S., Streeter S., Некрестьянов И.С., Соловьев А.Н.)
- Вероятностные алгоритмы (Minsky M., Воронцов К.В.)
- Эволюционные алгоритмы (Fraser A., Курейчик В.В.)

Основные модели представления знаний

- Терм-документная матрица (Landauer T., Deerwester S., Streeter S., Некрестьянов И.С., Соловьев А.Н.)
- Векторная модель (Salton G., Моченов С.В., Бледнов А.М., Луговских Ю.А.)
- Семантические БД (Miller G., Fellbaum C., Лукашевич Н.В., Добров Б.В. и др.)

Постановка задачи

- Имеется подборка текстовых данных (товары, услуги)
- Имеется подборка данных пользователей (покупателей, поставщиков)
- Данные распределены неравномерно между категориями
- Необходимо обработать данные так, чтобы их можно было использовать для подбора персональных рекомендаций для любого пользователя

Алгоритм

- Подготовка данных (для всех документов)
  - очистка от стоп-слов
  - обработка стеммером Портера (переход от слов к термам)
  - определение вхождения терма в документ
- Получение набора термов (на основе обучающей выборки)
  - статистический анализ количества вхождений термов в документы, составление терм-документной матрицы
  - расчет матрицы корреспонденций термов (МКТ)
  - ортогональное разложение МКТ, выделение семантического ядра - отбрасывание малозначащих термов

- Построение категориальных векторов
  - обучение - получение списка категорий (на основе обучающей выборки)
  - расчет векторных моделей категорий в пространстве термов
  - построение категориальных векторов документов базы
- Подбор вакансий
  - расчет категориального вектора пользователя, для которого происходит подбор рекомендаций
  - расчет коэффициентов близости с категориальными векторами базы вакансий
  - сортировка по убыванию, извлечение q первых элементов

Подготовка данных к анализу

- Удаление стоп-слов
- Стемминг Портера
- Выделение семантического ядра

Стемминг - это процесс нахождения основы слова для заданного исходного слова.
Семантическое ядро - это подборка понятий, имеющих существенное значение для данной предметной области.

Составление терм-документной матрицы

Рассмотрим матрицу X с n столбцами $\vec{x_i}$, где n - количество термов, и вектора термов равны:

$$x_j = \{x_{1j}, x_{2j}, \ldots, x_{mj}\}$$

где x_{ij} — частота встречаемости терма в документе:

$$x_{ij} = tf(t_j, d_i)$$

где d_i – i-ый документ из обучающей выборки, $i = 1, \ldots, m$, $tf(t_j, d_i)$ частота встречаемости терма t_j в документе d_i (term frequency). Матрица X называется *терм-документной матрицей*.

Матрица корреспонденций термов

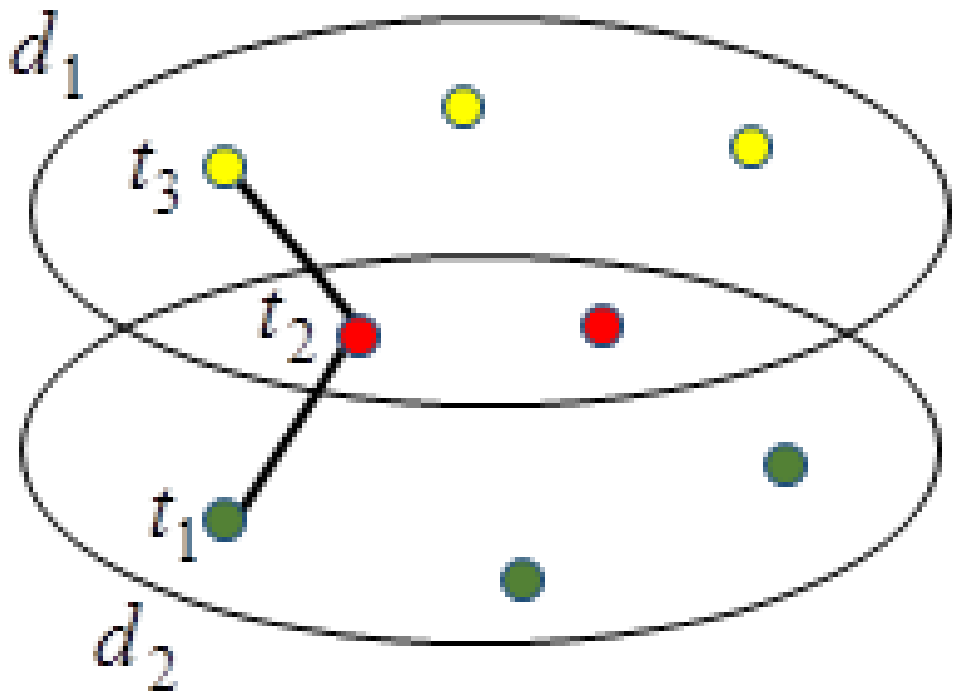
Определение. Матрицей корреспонденций термов $G = \{g_{ij}\}$ будем называть квадратную матрицу, элементами которой являются коэффициенты g_{ij}, отражающие близость i-го и j-го термов. Для элементов матрицы G выполняются условия:

1. $g_{ij} = g_{ji}$
2. $g_{ij} = 0$ при отсутствии взаимосвязи между термами.

Матрица G отображает взаимосвязи термов внутри документов на основе знаний о частоте их совместного употребления.

Иллюстрация взаимосвязей термов

На рисунке 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.



Расчет МКТ

Возьмем в качестве МКТ матрицу, полученную из произведений нормированных векторов-термов

$$G = \{(\vec{y}_i, \vec{y}_j)\}_{i,j=1}^{n} = Y^T Y, \tag{1}$$

где через y_j обозначен вектор-столбец $y_j = \{y_{1j}, y_{2j}, \ldots, y_{mj}\}$.

$$y_{ij} = \frac{x_{ij}}{\sum\limits_j x_{ij}} = \frac{x_{ij}}{n_i} \tag{2}$$

n_i - общее количество слов в документе d_i.

Ортогональное разложение МКТ

Утверждение

Ортогональное разложение матрицы корреспонденций термов G, определенной по формуле (1), имеет вид:

$$G = VSV^T \tag{3}$$

где V – ортогональная матрица правых сингулярных векторов в разложении нормированной терм-документной матрицы Y, матрица S – диагональная матрица, на диагонали которой стоят $\lambda_i = \sigma_i^2$, σ_i - сингулярные коэффициенты разложения матрицы Y.

Сингулярное разложение ТДМ

Теорема Дж. Форсайта
Для любой вещественной $n \times n$ матрицы A существуют $n \times n$ матрицы U и V такие, что

$$U^T AV = E, \quad UU^T = E, \quad VV^T = E,$$

где E - диагональная матрица с элементами $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0$, r - ранг матрицы A.

Сингулярное разложение матрицы X используется в латентно-семантическом анализе (S. Deerwester, T. Landauer).

Основная идея применения разложения состоит в том, что термы, имеющие высокие сингулярные коэффициенты σ_i, являются значимыми, а термы с низкими коэффициентами можно отбросить.

Если бы МКТ считалась без нормировки, т.е. $G = X^T X$, где X – ТДМ, то результаты отбрасывания термов, соотвествующих малым собственным числам МКТ, и латентно-семантического анализа совпали бы.

При использовании нормированной МКТ $G = Y^T Y$ результаты различаются.

Свойства собственных чисел МКТ

Будем рассматривать коллекцию документов $D = \{d_1, d_2, ..., d_m\}$ и набор термов $T = \{t_1, t_2, ..., t_n\}$. Будем предполагать, что длина первых документов k, где $1 \leq k \leq n$, равна Φ, а длина оставшихся ϕ. В этом случае терм-документная матрица $X = (x_{ij})_{i=1,j=1}^{m,n}$ запишется в виде

$$x_{ij} = \begin{cases} \Phi a_{ij} & \text{для} \quad i = \overline{1,k}, \quad j = \overline{1,n}, \\ \phi b_{ij} & \text{для} \quad i = \overline{k+1,m}, \quad j = \overline{1,n}. \end{cases}$$

Относительно чисел $\Phi, \phi, a_{ij}, b_{ij}$ будем предполагать выполненными следующие условия

$$a_{ij} \geq 0 \qquad \forall i = \overline{1,k}, \quad j = \overline{1,n}, \tag{4}$$

$$b_{ij} \geq 0 \qquad \forall i = \overline{k+1,m}, \quad j = \overline{1,n}, \tag{5}$$

$$\sum_{j=1}^{n} a_{ij} = 1 \qquad \forall i = \overline{1,k}, \tag{6}$$

$$\sum_{j=1}^{n} b_{ij} = 1 \qquad \forall i = \overline{k+1,m}, \tag{7}$$

$$\Phi > \phi \geq 1. \tag{8}$$

Матрицу X удобно записать в виде

$$X = \Phi A + \phi B, \tag{9}$$

где A - матрица $m \times n$, у которой элементами первых n строк являются элементы a_{ij}, а все элементы остальных $m - k$ строк равны нулю, B - матрица $m \times n$, у которой элементы первых k строк равны нулю, а элементами остальных $m - k$ строк являются числа b_{ij}.

Вопрос: какое влияние на собственные числа матрицы G оказывают числа Φ и ϕ при условии, что Φ существенно больше чем ϕ?

Теорема

Пусть выполнены условия (4)–(8). Тогда для любого $s = \overline{1, n}$ *справедлива оценка*

$$\Phi^2 \lambda_s(G_A) \leq \lambda_s(G) \leq \Phi^2 \lambda_s(G_A) + \phi^2(m - k). \quad (10)$$

Влияние длины документа

Утверждение

Пусть длина всех документов в коллекции одинакова, т.е.

$$\sum_{j=1}^{n} n_{ij} = n_i = \Phi,$$

тогда $X = \Phi Y$ *и сингулярное разложение матрицы* X *имеет вид* $X = U(\Phi S)V^T$, *где унитарные матрицы* U *и* V *- матрицы в сингулярном разложении нормированной терм-документной матрицы, т.е.* $Y = USV^T$.

Теорема

Пусть проводится выделение семантического ядра из коллекции k длинных документов, длиной Φ, $m - k$ коротких документов, длиной ϕ, причем длина коротких документов удовлетворяет условию $\frac{\phi}{\Phi} < \epsilon$.
Тогда сингулярные числа $\sigma_s(X)$ в разложении терм-документной матрицы X близки к сингулярным числам $\sigma_s(\Phi A)$ терм-документной матрицы ΦA, содержащей только длинные документы, при $s = \overline{1, r_A}$, где r_A – ранг матрицы A.

Для сингулярных чисел матрицы X выполняется неравенство

$$\sigma_s(\Phi A) \leq \sigma_s(X) \leq \sigma_s(\Phi A)\left(1 + 0.5\epsilon^2 \cdot \frac{(m-k)}{\sigma_s^2(A)}\right) \tag{11}$$

если $s \leq r_A$. Здесь $\sigma_s(A)$ — s-е сингулярное число матрицы A, $\Phi\sigma_s(A) = \sigma_s(\Phi A)$.
При $s > r_A$ сингулярные числа матрицы X удовлетворяют неравенству

$$0 \leq \sigma_s(X) \leq \phi\sqrt{m-k}. \tag{12}$$

Пример

Пусть коллекция документов состоит из 4-х длинных документов, длина каждого из которых $n_i = \Phi = 500$ термов $(i = 1, \ldots, 4)$, и 2-х коротких, содержащих по $n_i = \varphi = 50$ термов $(i = 5, 6)$. Терм-документная матрица имеет вид

$$X = \begin{pmatrix} 285 & 60 & 90 & 55 & 10 \\ 105 & 30 & 265 & 80 & 20 \\ 117 & 25 & 167 & 151 & 40 \\ 82 & 48 & 155 & 132 & 83 \\ 14 & 16 & 2 & 8 & 10 \\ 4 & 18 & 12 & 10 & 6 \end{pmatrix}$$

Сингулярные коэффициенты матрицы X (без нормировки) равны $\{509.5, 197.9, 98.4, 40.1, 13.1\}$.
Найдем сингулярное разложение матрицы ΦA $\{508.9, 197.7, 97.7, 37.1, 0\}$.
Найдем относительные разности δ_s, $s = \overline{1,4}$, где

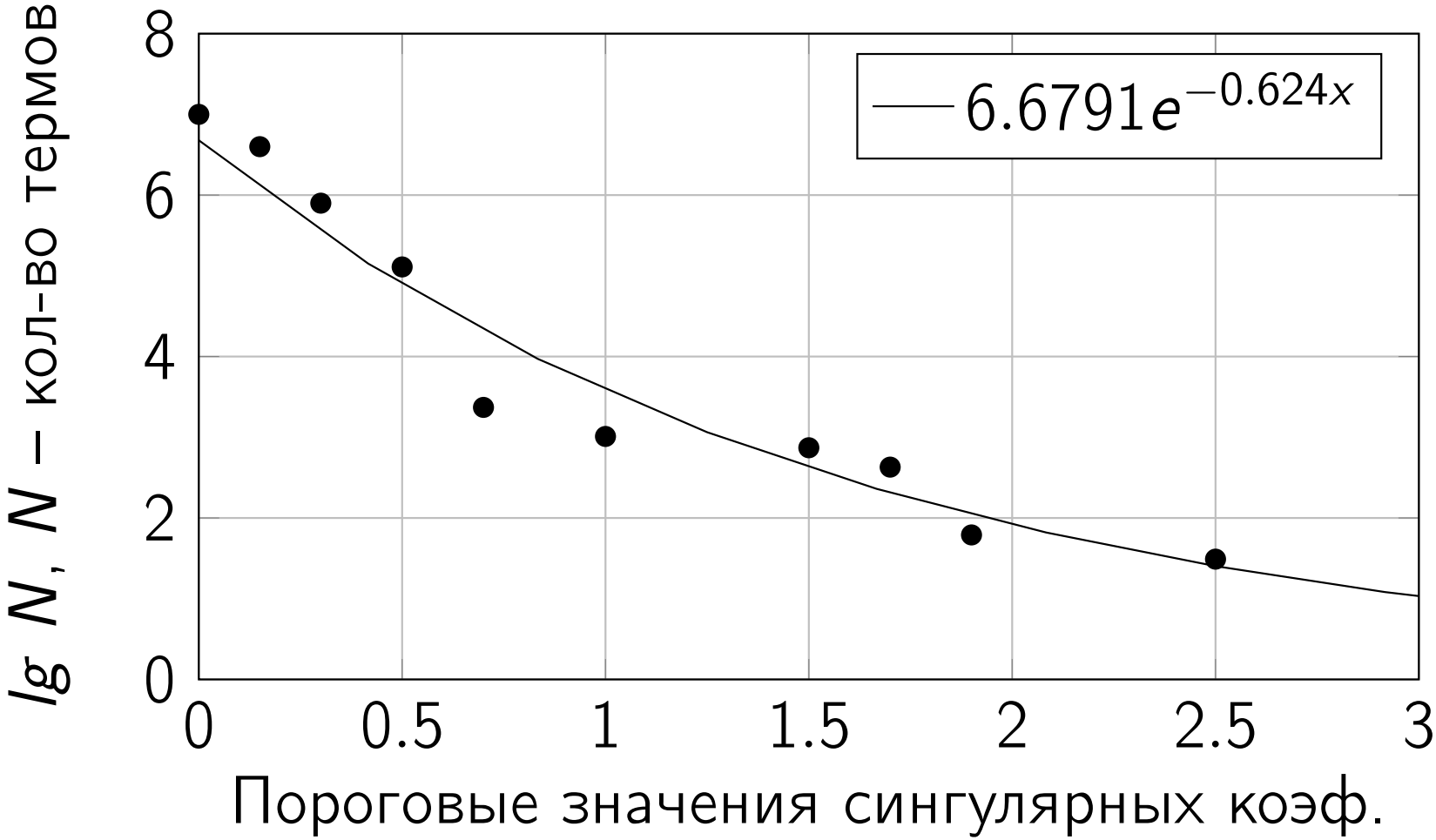
$$\delta_s = \frac{\sigma_s(X) - \sigma_s(\Phi A)}{\sigma_s(\Phi A)}.$$

Получим относительно небольшие значения $\{0.0012, 0.0011, 0.0077, 0.078\}$. Проверка показывает, что в данном примере неравенство (11) выполнено.

Обработка обучающей выборки (обучение)

- Выбор точности (размерности семантического пространства)
- Формирование списка категорий
- Отнесение каждого текста обучающей выборки к какой-либо категории

График зависимости



Уточнение понятия вхождения терма в текст

Пусть w_j - j-тое слово анализируемого текста d, t - текущий терм.
Считается, что t входит в D, если:

$$\exists w_j : (Ham(w_j, t) < h) \vee (stem(w_j) = stem(t))$$

где *stem* - операция стемминга, *Ham* - операция получения расстояния Хэмминга, h - расстояние, при котором можно считать слова идентичными.

Вычисление векторной модели категорий

Найдем средний вектор j-ой категории, в которую входит p текстов:

$$\vec{d}_{\text{ср.}j} = \left\{ \frac{\sum_{i=1}^{p} tf(t_1, d_i)}{p}, \ldots, \frac{\sum_{i=1}^{p} tf(t_p, d_i)}{p} \right\}$$

где d_i- i-тый документ из j-ой категории.

Построение категориальных векторов документов базы

Например, необходимо вычислить категориальный вектор для выбранного документа d. Обозначим его Z, тогда j-ая компонента будет иметь вид:

$$z_j = \frac{\left(\vec{d}_{\text{ср.}j}, \vec{d}\right)}{\left|\vec{d}_{\text{ср.}j}\right| \left|\vec{d}\right|}$$

где $\vec{d}_{\text{ср.}j}$ - средний вектор j-ой категории, $\vec{d}$ - вектор документа из коллекции.

Выбор рекомендаций

1. Рассчитывается категориальный вектор пользователя. Обозначим его $\vec{Z}_{\text{польз}}$
2. Будем называть γ_i коэффициентом близости между $\vec{Z}_{\text{польз}}$ и $\vec{Z}_i$, $\vec{Z}_i \in R^c$: $\gamma_i = \frac{\left(\vec{Z}_{\text{польз}}, \vec{Z}_i\right)}{\left|\vec{Z}_{\text{польз}}\right|\left|\vec{Z}_i\right|}$
3. Полученные значения сортируются по убыванию.
4. Из отсортированной выборки извлекается q первых элементов.

Мотивация использования семантической близости

- Решение проблемы синонимии и полисемии
- Значительное снижение размерности семантического пространства
- Легкость восприятия экспертом в предметной области

Известные способы

- Использование семантических БД (WordNET) (Лукашевич Н.В., Добров Б.В., Сухоногов А.М., Яблонский С.А.)
- Использование семантической близости (Крюков К.В., Панкова Л.А., Пронина В.А., Суховеров В.С.)
- Методы основанные на анализе веб-энциклопедий (Варламов М.В.)

Семантическая близость

- Некоторые пары слов по смыслу ближе друг к другу
  - Машина – шина семантически связаны друг с другом
  - Машина – дерево не связаны
- Семантическая близость между словами может состоять из:
  - Синонимии (машина – автомобиль)
  - Связи часть-целое (машина – шина)
  - Совместной встречаемости (машина – дорога)

Новый вес терма, использующий семантическую близость

$$\tilde{w}_{dt_1} = w_{dt_1} + \sum_{t_1 \neq t_2} similarity(t_1, t_2)$$

где w_{dt_1} - вес терма в документе d до настройки, рассчитанный по формуле, *similarity* - семантическая близость термов t_1 и t_2, рассчитываемая с помощью адаптации расширенного метода Леска. Суммирование происходит по всем термам документа d.

Семантическая близость с помощью метода Леска

$$similarity_{extLesk}(t_1, t_2) = overlap(gloss(t_1), gloss(t_2)+$$
$$overlap(gloss(hyppo(t_1)), gloss(hyypp(t_2))+$$
$$overlap(gloss(hyppo(t_1)), gloss(t_2))+$$
$$overlap(gloss(t_1), gloss(hyppo(t_2)))$$

где $overlap(t_1, t_2)$ - количество совпадений между термами t_1 и t_2, $gloss(t)$ - определение терма t, $hyppo(t)$ - гипероним слова, например для слова «красный» гиперонимом является слово «цвет», t_1 и t_2 - термины.

Способы оценки

- $F\text{-}measure = 2 \cdot \frac{precision \cdot recall}{precision + recall}$
  где *precision* - количество правильных результатов в выдаче алгоритма, *recall* - общее количество результатов выдачи.
- $purity(W, C) = \sum_k \max_j |w_k \cup c_j|$ где W - множество документов, w_k - k-тый документ, C - множество категорий - множество документов, отнесенных классификатором к категории k, c_j - множество документов, отнесенных к категории j экспертом.

Сведения о выборках и способы оценки

- Объявления о поиске работы 700 тыс.
- Новости 1.2 млн.
- Литературные аннотации 20 тыс.

	Алгоритм из работы		Алгоритм с использованием семантической близости	
Множество	F-measure	Purity	F-measure	Purity
Объявления о работе	0.31	0.33	0.65	0.66
Новости	0.56	0.58	0.61	0.64
Литературные аннотации	0.56	0.57	0.63	0.67

Проблемы современных подходов к вычислению семантической близости

- "Теннисная"проблема в WordNET
- Охват только общей лексики и семантики естественного языка
- Сложность учета синонимии и полисемии при использовании алгоритмах, основанных на анализе веб-источников

Полисемия

Вычислим связь между близкими словами и совершенно различными словами, используя «синсеты» WordNet и метод Леска и метод Резника. Результаты приведены в таблице.

Таблица: Зачения близости

Метод расчета близости	Леска	Резника
«университет» и «экзамен»	0.2	0
«университет» и «растение»	0.28	0.23

Сравнивая результаты, можно прийти к выводу, что слово «университет» ближе к слову «растение», чем к слову «экзамен».

Контекстное множество терма

Определение. Будем называть множество слов, связанных с заданным термом *контекстным множеством* терма.
Для построения контекстного множества будем использовать матрицу корреспонденций термов G, полученную из нормированной терм-документной матрицы:

$$G = ((y_i, y_j))_{i,j=1}^{n} = Y^T Y.$$

Предположим, необходимо построить контекстное множество i-того терма.

- Возьмем i-тую строку матрицы корреспонденций термов G.
- Вычислим среднее арифметическое среди элементов вектора. Обозначим среднее $\bar{G}_i$.
- Отбросим все компоненты вектора G_i меньшие среднего значения $\bar{G}_i$. Оставшиеся термы составят контектное множество i-того терма.

Вычисление семантической близости на основе контекстного множества

- Формирование контекстных множеств термов t_1 и t_2.
  Пусть $C_1 = \{c_{11}, c_{12}, ..., c_{1n}\}$ и $C_2 = \{c_{21}, c_{22}, ..., c_{2m}\}$ - контекстные множества термов t_1 и t_2 соответственно. Сформируем общее контекстное множество: $C = C_1 \cup C_2$. Мощность данного множества не больше $n + m$.

- Вычисление нормализованных близостей между общим контестным множеством и каждым из термов t_1 и t_2:

$$близость(c_i, t_1) = \frac{частота(c_i, t_1)}{макс.частота(t_1)},$$

$$близость(c_i, t_2) = \frac{частота(c_i, t_2)}{макс.частота(t_2)},$$

где $частота(c_i, t_1)$ - количество документов, где c_i и t_1 встречаются вместе.

макс.частота(t_j) рассчитывается по формуле:

$$\textit{макс.частота}(t_j) = \max\left\{\textit{частота}(c_i, t_j)\right\}, c_i \in C$$

- Расчет семантической близости
  Рассчитаем коэффициенты R_i для всех слов из контекстного множества C по формуле:

$$R_i = \frac{\min\left\{\textit{близость}(c_i, t_1), \textit{близость}(c_i, t_2)\right\}}{\max\left\{\textit{близость}(c_i, t_1), \textit{близость}(c_i, t_2)\right\}}$$

$$\textit{сем.близость}(w_1, w_2) = \frac{\sum_{i=1}^{k} \left(\frac{p_i R_i}{1+R_i} + s \right)}{1+s}$$

где $p_i = 2$, когда оба терма встречаются в одном документе,

$p_i = 1$ противном случае,

s - коэффициент синонимии,

$s = 1$, если слова t_1 и t_2 являются синонимами,

$s = 0$ в противном случае.

Основные результаты

- Способ формирования семантического пространства на основе матрицы корреспонденций термов, которая подвергается ортогональному разложению для компактного отображения текста в семантическое пространство.
- Использование вычисления категориальных векторов для упорядочения результирующей выборки по степени релевантности запросу пользователя для гарантии получения непустого результата, независимо от распределения обучающей выборки.
- Метод перевзвешивания термов векторной модели с помощью вычисления их семантической взаимосвязи друг с другом на основе авторской адаптации алгоритма Леска.
- Статистический метод вычисления семантической близости термов, основанный на вычислении специально подбираемых контекстных множеств термов.

Практическое применение

В работе разработан комплекс методов интеллектуального анализа данных и моделей представления знаний, направленных на применение знаний о семантической и лексикографической взаимосвязи слов для улучшения качества приложений информационного поиска, систем поддержки принятия решений, систем подбора персональных рекомендаций и т.п. Разработанные подходы позволили значительно снизить объем хранимых промежуточных данных.

Статьи в журналах их перечня ВАК

- Бондарчук Д.В. Статистический способ определения семантической близости термов // Системы управления и информационные технологии. – 2015. – Т. 61, № 3. – С. 55–57.
- Бондарчук Д.В. Алгоритм построения семантического ядра для текстового классификатора // В мире научных открытий. – 2015. – Т. 68, № 8.2. – С. 713–724.
- Бондарчук Д.В., Тимофеева Г.А. Выделение семантического ядра на основе матрицы корреспонденций термов // Системы управления и информационные технологии. – 2015. – Т. 61, № 3.1. – С. 134–139.
- Бондарчук Д.В., Тимофеева Г.А. Применение машинного обучения для формирования персональных рекомендаций в сфере трудоустройства // Экономика и менеджмент систем управления. – 2015. – Т. 18, № 4.2. – С. 215—221.
- Бондарчук Д.В., Тимофеева Г.А. Математические основы метода категориальных векторов в интеллектуальном анализе данных // Вестник Уральского государственного университета путей сообщения. – 2015. – 4(28). – С. 4–8.

Статья в издании, индексируемом в Web of Science

- Bondarchuk D.V., Timofeeva G.A. Vector space model based on semantic relatedness // Proceedings of International Conference «Applications of Mathematics in Engineering and Economics» (AMEE'15) . | 2015. | DOI: 10.1063/1.4936683.

Статьи в изданиях, индексируемых в РИНЦ

- Бондарчук Д.В. Использование латентно-семантического анализа в задачах классификации текстов по эмоциональной окраске // Бюллетень результатов научных исследований. – 2012. – 2(3). – С. 146–151.
- Бондарчук Д.В. Выбор оптимального метода интеллектуального анализа данных для подбора вакансий // Информационные технологии моделирования и управления. – 2013. – 6(84). – С. 504–513.
- Бондарчук Д.В. Интеллектуальный метод подбора персональных рекомендаций, гарантирующий получение непустого результата // Информационные технологии моделирования и управления. – 2015. – Т. 2(92).–С. 130–138.