Приложения на линейната алгебра в методи за извличане на информация (text mining).

Векторен модел за извличане на информация

Извличане на информация (Information retrieval, IR)

- Hans Peter Luhn (1896–1964) пръв предлага статистически модел за извличане на информация, базиран на понятието "сходство" (1957 г.).
- През 60-те години Gerard Salton и колегите му започват разработването на **векторния модел за извличане на информация** от масив от текстови документи. Документите и потребителската заявка за търсене се представят като вектори, като броят на координатите им (размерността на пространството) е равен на броя на думите (terms) (които могат да участват в търсенето). Векторите могат да формират матрица (term-document matrix).
- В края на 80-те години се появяват и други методи, базирани на представяния на тази матрица с цел намаляване на нейния ранг: латентно семантично индексиране (Latent Semantic Indexing, LSI), създаден от Susan Dumais (1988), който използва разлагане по особени стойности (Singular value decomposition, SVD); неотрицателно матрично факторизиране (Nonnegative Matrix Factorization) и др.

Векторен модел за извличане на информация

Текстовото съдържание на всеки документ (файл) се представя като **вектор** (наредена n-торка), където n е броят на думите, участващи в търсенето. Всяка координата на векторите е неотрицателно число и трябва да отразява значимостта на съответната дума в конкретния документ и в целия масив.

Два въпроса:

- Какво да бъде **числовото представяне** на съдържанието на всеки документ, т.е. как да се определят стойностите на **координатите на векторите**?
  Бинарно (0 и 1, най-простото представяне), брой повторения на съответната дума (tf = term frequency), TF-IDF метод (term frequency inversed document frequency) и др.
- Как да бъдат **сравнявани за сходство** (или различие) векторите, съответстващи на документите и потребителската заявка за търсене?
  Мерки за сходство (подобие) (Similarity measures) и мерки за различие (разстояние) (Distance or Dissimilarity measures).

Методи за определяне на координатите на векторите

Бинарният метод е първият използван и най-простият модел. При този подход на всеки документ (и на потребителската заявка за търсене) се съпоставя вектор с координати x_i, равни само на нули или единици, като

$$x_i = \begin{cases} 1, \text{ ако } i\text{-тата дума се съдържа в текста на документа,} \\ 0, \text{ ако } i\text{-тата дума не се съдържа в текста на документа.} \end{cases}$$

Не се отчита броят на повторенията на думите в документите.
Получените вектори, съответстващи на документите, могат да бъдат подредени по стълбовете (или редовете) на матрица (term-document matrix). Това важи и за останалите методи.

Методи за определяне на координатите на векторите

Отчитане на броя на повторенията (term frequency, tf) на всяка дума в текстовете на документите. При този подход i-тата координата на вектора на документа е равна на броя на повторенията на думата tf_i в текста на съответния документ.

При търсене с участието на i-тата дума, документ, за който $tf_i = 10$, е по-релевантен (значим) от такъв, за който $tf_i = 1$. Но не е 10 пъти по-значим, т.е. релевантността е пропорционална на броя на повторенията, но не нараства с коефициент на прорционалност, равен на броя на повторенията.

Затова вместо tf_i може да се използва логаритмична **теглова функция** wf_i за i-тата дума, която се определя от

$$wf_i = \begin{cases} 1 + \log_{10} tf_i, & \text{ако } tf_i \geq 1, \\ 0, & \text{ако } tf_i = 0. \end{cases}$$

Съществуват и други варианти за теглови функции.

Методи за определяне на координатите на векторите

Съгласно H. P. Luhn, думи, които се срещат много често, както и думи, които се срещат много рядко в масив от сходни по съдържание документи, имат малък ефект върху подреждането по релевантност на документите.

Ако разглеждаме масив от тематично сходни документи, колко на брой често повтарящи се и колко рядко повтарящи се думи трябва да очакваме да открием?

Отговор на горния въпрос дава известният **Закон на Зиф** (George Kingsley Zipf – американски лингвист, изследвал статистически зависимости в естествени езици). Той установява, че в естествените езици има малко на брой често повтарящи се думи и много на брой рядко повтарящи се. Законът на Зип гласи, че i-тата най-често срещана дума в даден текст има честота на срещане, пропорционална на $\frac{1}{i}$, т.е. втората най-често срещана дума се повтаря два пъти по-малко от първата най-срещана, третата – три пъти по-малко от първата и т.н.

Методи за определяне на координатите на векторите

Чрез величината **tf** отчитаме значимостта на всяка дума за съответния документ (чрез броя на повторенията ѝ в текста или чрез претегления брой повторения).
Чрез величината **idf** (inversed document frequency) отчитаме значимостта на всяка дума за целия масив от документи, като взимаме предвид броя на документите, съдържащи думата (с обратно пропорционална зависимост). За целта най-често се използва формулата (Spärck-Jones, 1972; 2004)

$$idf_i = \log_{10}\left(\frac{N}{n_i}\right),$$

където N е броят на всички документи в масива, а n_i е броят на документите, съдържащи i-тата поредна дума.

Теглото idf има по-голям ефект при търсения, в които участват две или повече думи и малък ефект при търсения от една дума.

Методи за определяне на координатите на векторите

Класическите **TF-IFD метод или WF-IDF метод** (Salton, Wong, Yang, 1975; Salton, 1983; Salton, Buckley, 1987)

Най-често използваните методи за определяне на координатите на векторите. Базират се на комбиниране чрез произведение на tf или wf с idf съответно съгласно формулите:

$$tf_i \times idf_i = tf_i \log_{10}\left(\frac{N}{n_i}\right), \qquad wf_i \times idf_i = (1+\log_{10} tf_i)\log_{10}\left(\frac{N}{n_i}\right).$$

Произведенията имат най-големи стойности за думи, които се повтарят много на брой пъти в малко на брой документи. По-ниски стойности се получават за думи, които се повтарят малко на брой пъти в даден документ, или се срещат в голяма част от документите. Най-ниски стойности се получават за думи, които се срещат във всички документи от масива.

Варианти на TF-IDF: http://www.minerazzi.com/tutorials/term-vector-3.pdf.

Разстояние и подобие – математически определения

Нека V е множество. Функцията $d : V \times V \to \mathbb{R}$ се нарича **разстояние** (различие) върху V, ако за всеки $x, y \in V$ са изпълнени свойствата:

- $d(x, y) \geq 0$ (неотрицателност),
- $d(x, y) = d(y, x)$ (симетричност),
- $d(x, x) = 0$.

Функцията $s : V \times V \to \mathbb{R}$ се нарича **подобие** (сходство) върху V, ако:

- $s(x, y) \geq 0$ (неотрицателност),
- $s(x, y) = s(y, x)$ (симетричност),
- $s(x, y) \leq s(x, x)$

за всеки $x, y \in V$, като равенство в последното условие се достига само при $x = y$.

Разстояние d от сходство $s \in (0; 1]$ се получава чрез: $d = 1 - s$, $d = \frac{1-s}{s}$, $d = \sqrt{1 - s}$, $d = \sqrt{2(1 - s^2)}$, $d = -\log_{10} s$ и др.

Най-често използвани мерки на сходство в IR

Нека $x = (x_1, x_2, ..., x_n)$ и $y = (y_1, y_2, ..., y_n)$ са вектори с еднаква размерност.

- **Косинусът на ъгъла** между x и y (cosine similarity) е най-често използваният показател за подобие

$$\cos(x, y) = \frac{xy}{||x||.||y||},$$

  където $xy = x_1y_1 + x_2y_2 + ... + x_ny_n$ е скаларното произведение на x и y, а $||x|| = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$ е евклидовата дължина (норма) на x.

- **Коефициент на подобие на Жакард** (Jaccard similarity)

$$J(x, y) = \frac{xy}{||x||^2 + ||y||^2 - xy}.$$

- **Коефициент на подобие на Сьоренсен-Дайс** (Dice similarity)

$$D(x, y) = \frac{2xy}{||x||^2 + ||y||^2}.$$

Разстояния в IR

Разстоянието на Минковски (p-разстояние) между векторите $x = (x_1, x_2, ..., x_n)$ и $y = (y_1, y_2, ..., y_n)$ се определя като p-нормата на разликата $x - y$, т.е. съгласно формулата

$$||x - y||_p = (|x_1 - y_1|^p + |x_2 - y_2|^p + ... + |x_n - y_n|^p)^{\frac{1}{p}},$$

където $p \geq 1$.

При $p = 2$ се получава евклидовото разстояние между x и y (евклидовата норма на $x - y$). Евклидовото разстояние не е подходящо за прилагане върху вектори с различни дължини. Ако x и y са с равни дължини (напр. за нормирани вектори, т.е. $||x|| = ||y|| = 1$), дава същите резултати като косинуса на ъгъла между x и y. При $p = 1$ се получава разстояние по единична норма (Manhattan distance)

$$||x - y||_1 = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n|,$$

а при $p \to \infty$ се получава разстояние по безкрайна норма

$$||x - y||_\infty = \max_{i=1,...,n} |x_i - y_i|.$$

Разстояния в IR

Може да се използват и т. нар. претеглени разстояния на Минковски

$$||x - y||_p = \left(\sum_{i=1}^{n} w_i |x_i - y_i|^p \right)^{\frac{1}{p}},$$

където числата w_i се наричат тегла и придават различна "значимост" на различните координати. Използва се също и за анализ на подобие на изображения (например при разпознаване на изображения).

Предимства и ограничения на векторния модел

Основни предимства:

- опростен модел, базиран на линейна алгебра;
- сходствата могат да приемат всички стойности в $[0; 1]$, а не само отделни дискретни стойности;
- позволява подреждане на документите съгласно тяхната релевантност чрез различни показатели за сходство/различие.

Основни ограничения:

- проблем създава т. нар. "лингвистичен шум" – синоними и думи с повече от едно значение; ключовите думи в търсенето трябва да съвпадат точно с думите в документите;
- при формирането на векторите не се отчита редът, в който думите се срещат в текста;
- теоретично се предполага, че думите са статистически независими.

Литература

- C. D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval, Cambridge University Press, 2008.
- A. N. Langville, C. D. Meyer. Information Retrieval and Web Search, In: Handbook of Linear Algebra, 2nd ed., edited by Leslie Hogben, CRC Press, 2014.
- S. Dominich. The Modern Algebra of Information Retrieval, Springer, 2008.
- L. Eldén. Matrix Methods in Data Mining and Pattern Recognition, SIAM, 2007.
- S. Büttcher, C. L. Clarke, G. V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines, MIT Press, 2010.
- M. W. Berry, M. Browne. Understanding Search Engines: Mathematical Modeling and Text Retrieval (Software, Environments, Tools), 2nd ed., SIAM, 2005.
- M. M. Deza, E. Deza. Encyclopedia of Distances, 4th ed., Springer, 2016.

Литература

- D. Dublin. The Most Influential Paper Gerard Salton Never Wrote.
- D. L. Lee, H. Chuang, K. Seamons. Document Ranking and the Vector-Space Model, 1997.
- E. Garcia. The Classic TF-IDF Vector Space Model.