TEXT BASED DOCUMENT SIMILARITY MEASURE

Shnibekov Zhasulan

ABSTRACT

Do you have a shortage of data? Not very likely. A consequence of the pervasive use of computers is that most data originate in digital form. If we trade a stock or write a book or buy a product online, these events evolve electronically. Since so many paper transactions are now in paperless digital form, lots of "big" data are available for further analysis.

The concept of data mining, finding valuable patterns in data, is an obvious response to the collection and storage of large volumes of data. Data mining is no longer an emerging technology awaiting further development. Although its application is far from universal, the techniques of data mining are highly developed and for some forms of analysis are entering a mature phase.

We would like to say "Give us data and we will find the patterns."

Unfortunately, data-mining methods expect a highly structured format for data, necessitating extensive data preparation. Either we have to transform the original data, or the data are supplied in a highly structured format.

Data-mining methods learn from samples of past experience. If we speak to specialists in predictive data mining, their data will be in numerical form. These people are the "numbers guys." The "text miners" do not expect an orderly series of numbers. They are happy to look at collections of documents, where the contents are readable and their meaning is obvious.

This is our first distinction between data and text mining: numbers versus text. That doesn't mean that these are two distinct concepts. Both are based on samples of past examples. The composition of the examples is very different, yet many of the learning methods are similar. That's because the text will be processed and transformed into a numerical representation.

АННОТАЦИЯ

В современном мире нет понятия дефицита данных и информаций. Большинство данных представлены в цифровом виде из-за повсеместного использования компьютеров и компьютерных технологий. Такие события, как покупка различных товаров в интернете, торговля различными акциями, публикация книги, происходят в электронном виде. Так как большинство бумажных сделок происходят в электронной форме, много "больших" данных нуждаются в анализе.

Концепция интеллектуального анализа данных - это поиск закономерностей, которые являются очевидным ответом на сбор и хранение больших объемов данных. Технологии анализа данных актуальны на сегодня и не нуждаются в доработке. Хотя применение далеко не универсально, однако методы интеллектуального анализа данных являются весьма развитыми. К

сожалению, для применения этих методов, необходимо чтобы данные имели хорошо структурированный формат. В большинстве случаях данные нуждаются в переводе в тот вид, который требуют методы анализа.

Методы анализа основаны в сравнении образцов данных. Специалисты работают с данными, которые представлены в численном виде. Но большинство документов имеют текстовый вид, содержание которых легко можно прочитать.

Методы анализа обрабатывают численные значения. Однако текст можно легко представить в цифровом виде. Это единственная разница между текстом и числами в анализе данных.

ТҮЙІН

Қазіргі әлемде деректердің және информация тапшылығының ұғымы жоқ. Деректердің көпшілігі компьютердің және компьютерлік технологияның игерушілігінің мол болғаны соң, цифрлық көріністе ұсынады. Мысалы, интернеттен түрлі тауарларды сатып алған жағыдайда, түрлі акциялармен сауда жүргізгенде, кітаптың жарияланымы, электрондық көріністе болып жатады. Себебі қағаздық мәмілесінің көпшілігі құжаттандырылмаған соң, көп "кесек-кесек" деректер анализда мұқтаж болады.

Деректердің зияткерлік анализының тұжырымдамасы - ол бір заңдылықтың ізденісі. Деректер жиыны және деректердің кесек-кесек көлемінің сақтауы айқын жауап болып табылады. Деректер анализының технологиялары осы заманда өзектілік және пысықтауға мұқтаж емес. Қолданысы да ұзақ әмбебап емес, алайда деректердің зияткерлік анализының әдістері ең дамыған болып табылады. Өкінішке қарай, осы әдістерді қолдану үшін, деректер жөн құрылған форматта болуы тиіс. Көп жағыдайда, деректер анализының әдістері, мазмұнысы керек көрініске аударылуын мұқтаж етеді.

Анализ әдістерінің негіздері - деректердің үлгісінің салыстыруымы. Мамандар сандық көріністе ұсынған деректермен жұмыс істейтін. Бірақ құжаттың көпшілігі мәтіндік көріністе болған соң, мазмұнысы жеңіл оқылынады.

Анализдың әдістері сандық мағыналарды өңдейді. Алайда мәтінді цифрлық көрініске аударуы жеңіл болып табылады. Деректердің анализында, мәтіннің және сандардың айырылымашылығы осы.

FORMULATION OF A PROBLEM

Simple Desktop application to see the difference between two text based documents. Application will use various types of similarity measure functions. The idea of measuring text based document similarity has received considerable attention in several domains, including information retrieval and text mining. To begin with, we first transfer data into numerical vectors, and then we use similarity measure functions that would measure document similarities in general. Transferring data into numerical vectors

are tough work, where we use tokenization, filtering stopwords, stemming, etc, and then calculating by functions. In the end when data is in terms of numerical vectors then we use all the similarity measure functions.

Tokenization

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis.

Stopwords

In computing, stopwords are words which are filtered out prior to, or after, processing of natural language data (text). For some search machines, these are some of the most common, short function words, such as the, is, at, which and on. Default English stopwords list: a, about, above, after, again, against, all, am, an, and, any, are, …etc.

Stemming

In linguistic morphology, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

Vector generation

Vector generation (or *vector space model*) is an algebraic model for representing text documents as vectors of identifiers.

Documents and queries are represented as vectors.

$dj = (w_{1,j}, w_{2,j}, ..., w_{t,j})$

$q = (w_{1,q}, w_{2,q}, ..., w_{t,q})$

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero.

IDF-*inverse document frequency* is a measure of the general importance of the term

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$|D|$: the total number of documents

$|\{d : t_i \in d\}|$ number of documents where the term t_i appears (that is).

If the term is not in, this will lead to a division-by-zero:

$$1 + |\{d : t_i \in d\}|$$

TF-*term frequency*, defined as:

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j, that is, the size of the document $| d_j |$.

Finally,

$$(\text{tf-idf})_{i,j} = \text{tf}_{i,j} \times \text{idf}_i$$

Experimental analysis

Consider a document containing 100 words wherein the word *cow* appears 3 times. The term frequency (TF) for *cow* is then (3 / 100) = 0.03. So, TF = 0.03.

Now, assume we have 10 million documents and *cow* appears in one thousand of these. Then, the inverse document frequency is calculated as log(10 000 000 / 1 000) = 4. IDF = 4.

The TF-IDF score is the product of these quantities: 0.03 × 4 = 0.12.

TF-IDF = 0.12

Example 2:

Suppose the query "gold silver truck". The database collection consists of three documents (D = 3) with the following content

D1: "Shipment of gold damaged in a fire"

D2: "Delivery of silver arrived in a silver truck"

D3: "Shipment of gold arrived in a truck"

Table 1 shows us the result of calculation.

Table 1:

Term vector model based on w(i) = tf(i)*IDF(i)											
Query, Q: “gold silver truck” D1: "Shipment of gold damaged in a fire" D2: "Delivery of silver arrived in a silver truck" D3: "Shipment of gold arrived in a truck" D = 3; IDF = log(D/df(i))											
		Counts, tf(i)						**Weights**, w(i) = tf(i)*IDF(i)			
TERMS		**1**	**2**	**3**	**df(i)**	**D/df(i)**	**IDF(i)**	**Q**	**D1**	**D2**	**D3**
a					3	3/3=1	0	0	0	0	0
arrived					2	3/2=1.5	0.1761	0	0	0.1761	0.1761
damaged					1	3/1=3	0.4771	0	0.4771	0	0
deliver					1	3/1=3	0.4771	0	0	0.4771	0

y											
fire					1	3/1=3	0.4771	0	0.4771	0	0
gold					2	3/2=1.5	0.1761	0.1761	0.1761	0	0.1761
in					3	3/3=1	0	0	0	0	0
of					3	3/3=1	0	0	0	0	0
silver					1	3/1=3	0.4771	0.4771	0	0.9542	0
shipment					2	3/2=1.5	0.1761	0	0.1761	0	0.1761
truck					2	3/2=1.5	0.1761	0.1761	0	0.1761	0.1761

Similarity Analysis

First for each document and query, we compute all vector lengths (zero terms ignored)

$$|D1| = \sqrt{0.4771^2 + 0.4771^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.5173} = 0.7192$$

$$|D2| = \sqrt{0.1761^2 + 0.4771^2 + 0.9542^2 + 0.1761^2} = \sqrt{1.2001} = 1.0955$$

$$|D3| = \sqrt{0.1761^2 + 0.1761^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.1240} = 0.3522$$

$|Di| = \sqrt{\sum_i w^2 i,j}$

$$|Q| = \sqrt{0.1761^2 + 0.4771^2 + 0.1761^2} = \sqrt{0.2896} = 0.5382$$

$|Q| = \sqrt{\sum_i w^2 Q,j}$

Next, we compute all dot products (zero products ignored)

Q*D1 = 0.1761 * 0.1761 = 0.0310
Q*D2 = 0.4771 * 0.9542 + 0.1761 * 0.1761 = 0.4862
Q*D3= 0.1761 * 0.1761 + 0.1761 * 0.1761 = 0.0620

$$Q*Di = \sqrt{\sum_i w_{Q,j}\, w_{i,j}}$$

Now we calculate the similarity values

$$\text{Cosine}\theta_{D1} = \frac{Q*D1}{|Q|*|D1|} = \frac{0.0310}{0.5382*0.7192} = 0.0801$$

$$\text{Cosine}\theta_{D2} = \frac{Q * D2}{|Q| * |D2|} = \frac{0.4862}{0.5382 * 1.0955} = 0.08246$$

$$\text{Cosine}\theta_{D3} = \frac{Q * D3}{|Q| * |D3|} = \frac{0.0620}{0.5382 * 0.3522} = 0.3271$$

$\text{Cosine}\theta_{Di}$ = Sim (Q,Di)

$$\text{Sim (Q,Di)} = \frac{\sum_i w_{Q,j}\, w_{i,j}}{\sqrt{\sum_i w^2 Q,j}\,\sqrt{\sum_i w^2 i,j}}$$

Rank 1: Doc 2 = 0.8246
Rank 2: Doc 3 = 0.3271
Rank 3: Doc 1 = 0.0801

CONCLUSION

The application of these algorithms is the optimal solution for the analysis and comparison of text data.

REFERENCE

• [1] Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Fred J. Damerau, “Text mining: Predictive Methods for Analyzing Unstructured Information”, Springer, USA, 2005, pp. 1-2, 15-25, 85-89.

• [2] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, “Introduction to Information Retrieval”, Cambridge University Press, USA, 2008, pp. 117-119.

• [3] Leo Egghe, “New relations between similarity measures for vectors based on vector norms”,

• [4] Deza, Michel Marie, Deza, Elena, “Encyclopedia of Distances”, Springer, Berlin, 2009, pp.298-304.

УДК: 004.420

Online Ordering Taxi on Android

Makhymgaliyeva Gulden 4D_04

Nowadays people are very busy, stressful and always in an extremely active life. Sometimes they need relax, just sit and think about how life is wonderful. Also they need some service from another people, and to get agreement with someone when you are very tired is so difficult. This kind of problem makes you annoying. So I present decision of this issue by using mobile application Get a Taxi. This application locates the passenger's position automatically or can be set to pick-up from the user's favorite locations, e.g. work, home etc. The application then finds and orders the nearest available taxi and informs the user of the driver's name and ratings, and how much will cost distance. Map shows the passenger's position and the position of the taxi and displays the distance left and the estimated time of arrival. Booking and managing rides is quick and easy, saving you time and hassle. That is an awesome convenience when you are in a rush. A must-have friend in your pocket, ready for when you need it. The passenger can track the taxi‘s arrival on the map including time of arrival as well as the driver’s profile with picture, name, rating and phone number.

The word “mobile” means capable of changing quickly from one state or condition to another, tending to travel and change settlements frequently, e.g. "a highly mobile face". What about “mobile