Abstract

Information retrieval systems are very important as they permit us to easily access huge amounts of data. In the current context where we are having a lot of information, the precision of an information retrieval system is crucial. Besides retrieving content, having mechanisms that can compare different entities of text and group them by similarity is extremely important. A model based on time series can be used as a viable alternative to the well-known vector space model. The time series model takes into account the order in which the words appear and moreover it can be easily extended, in order to add semantics by defining a semantic function on individual words. The generic model that we are proposing can be easily integrated with different criteria in order to compute distances between text corpuses and thus to find text entities that are similar to each other. Given the unstructured nature of the natural language, this problem raises different challenges in terms of both precision (it is difficult to determine the similarity of two or more entities in the absence of structure) and performance (analyzing unstructured data requires complex algorithms that usually have a high time complexity). Therefore, the approach proposed in this work provides: a generic framework that makes use of time series, which can be used in information retrieval and natural language processing, improves the similar text search results quality and identifies texts that belongs to the same source.

Sumar

Sistemele de regăsire a informației sunt foarte importante, deoarece cu ajutorul lor putem accesa cantităţi mari de date. În momentul de față, avem acces la foarte multe date, motiv pentru care precizia unui astfel de sistem este esenţială. De asemenea, este important ca pe lângă găsirea unor informaţii să dispunem de un sistem capabil să compare două entităţi de text şi să le grupeze pe baza similitudinii. Astfel, un model bazat pe serii de timp poate reprezenta o alternativă la bine cunoscutul model bazat pe spaţii vectoriale. După cum vom vedea, modelul bazat pe serii de timp ţine cont de ordinea de apariţie a cuvintelor în propoziţie şi poate fi extins uşor, pentru a ţine cont şi de semantică – acest lucru se realizează uşor prin integrarea unei funcţii semantice de calcul a distanţei dintre cuvinte. Modelul generic pe care îl propunem poate fi uşor extins cu alte criterii care pot fi folosite pentru a calcula distanţa şi, respectiv, similitudinea dintre texte. Natura nestructurată a limbajului natural duce la apariţia diferitelor provocari, datorită cerinţelor de precizie (este dificil de determinat similitudinea dintre două sau mai multe documente în absenţa structurii) şi performanţa (analiza datelor nestructurate necesită algoritmi care, în general, au complexitate ridicată). Astfel, această lucrare propune: o arhitectură bazată pe serii de timp, care poate fi folosită atât pentru regăsirea informaţiilor, cât şi pentru prelucrarea limbajului natural, o metoda care îmbunătăţeşte rezultatele obţinute la calcularea textelor similare şi o metodă care grupează textele pe baza autorului.