Piotr Borkowski
Metody semantycznej kategoryzacji w zadaniach analizy dokumentów tekstowych
Rozprawa doktorska

Streszczenie

Niniejsza praca poświęcona jest zagadnieniu kategoryzacji dokumentów tekstowych, czyli przypisywaniu im listy słów lub fraz charakteryzujących kategorie, do których dokument tematycznie przynależy. Kategorie te pochodzą ze struktury hierarchicznej o dużej liczbie elementów.

Zagadnienie kategoryzacji jest istotne z praktycznego punktu widzenia, szczególnie tam, gdzie zachodzi potrzeba klasyfikacji dokumentów z użyciem dużych zbiorów etykiet o hierarchicznych zależnościach. Podejście takie może być alternatywą lub wsparciem dla już istniejących klasyfikatorów, szczególnie w przypadku potrzeby klasyfikacji danych, dla których nie dysponuje się odpowiednimi zbiorami uczącymi.

W pracy zaproponowany został algorytm semantycznej kategoryzacji dokumentów, który działa bez konieczności posiadania specjalnie przygotowanego zbioru uczącego. Zamiast niego używane są już istniejące zasoby w postaci taksonomii. Na jego podstawie opracowano algorytmy do rzutowania kategorii. Jeżeli dysponuje się ustalonym zbiorem kategorii, pozwalają one rzutować je na zadany zbiór. W przeciwnym razie, korzystając z algorytmu agregacji, można w sposób nienadzorowany rzutować wyniki kategoryzacji na podprzestrzeń właściwą dla zbioru dokumentów, unikając przy tym rozmycia semantycznego. Ponadto, opracowano algorytm do semantycznej klasyfikacji, który bazuje na metodzie kategoryzacji, i dobrze sprawdza się w przypadku tworzenia komitetów, jako dodatek do klasyfikatorów działających w oparciu o podejście typu „worka słów", szczególnie zaś dobrze sprawdza się w przypadku tekstów z luką semantyczną (dane pochodzące z różnych źródeł lub podlegające różnemu rozkładowi danych w zbiorze uczącym i testowym). Powstał również heterogeniczny komitet klasyfikatorów, który pozwala łączyć algorytm semantycznej kategoryzacji i znanych dotąd klasyfikatorów. Zaproponowano zmodyfikowane miary semantycznego podobieństwa konceptów i kategorii oraz miary oceny poprawności kategoryzacji dla nierozłącznych hierarchicznych klas.

W pracy wykazano skuteczność zaproponowanego algorytmu kategoryzacji z wykorzystaniem polskojęzycznej Wikipedii jako zasobu semantycznego oraz pokazano możliwość przeniesienia go na inne zasoby semantyczne, jak np. taksonomia medyczna MeSH.

Piotr Borkowski
Metody semantycznej kategoryzacji w zadaniach analizy dokumentów tekstowych
Rozprawa doktorska

Abstract

This work is devoted to the issue of categorizing text documents, that is, assigning them a list of words or phrases that characterize the categories to which the document thematically belongs. These categories come from a hierarchical structure with a large number of elements.

The issue of categorization is important from a practical point of view, especially where there is a need to classify documents using large sets of labels, with hierarchical dependencies. Such an approach may be an alternative or support for already existing classifiers, especially in the case of the need to classify data for which there is no adequate teaching set.

The paper proposes an algorithm for semantic categorization of documents that works without a specially prepared training set. Instead, it uses existing resources in the form of taxonomies. Based on it, algorithms for category projection were developed. If a set of categories is available, they allow them to be projected onto a given set. Otherwise, using the aggregation algorithm, you can in unsupervised way project the categorization results into the proper subspace for the set of documents, while avoiding semantic blurring. The algorithm for semantic classification has been developed, which is based on the categorization method, and works well in the case of classifier committees, as an addition to classifiers based on the "bag of word" approach, and in particular works well for texts with a semantic gap (data coming from different domains or subject to different data distribution in the training and test set). There was also created a heterogeneous classifiers committee, which allows combining the algorithm of semantic categorization and previously known classifiers. There were proposed modified measures of semantic similarity of concepts and categories as well as measures of correctness of categorization for inseparable hierarchical classes.

The paper shows the effectiveness of the proposed categorization algorithm using the Polish-language Wikipedia as a semantic resource and indicates the possibility of transferring it to other semantic resources, such as the medical taxonomy MeSH.