Università degli Studi di Salerno

Dipartimento di Ingegneria Elettronica ed Ingegneria Informatica

Dottorato di Ricerca in Ingegneria dell'Informazione
XI Ciclo – Nuova Serie

TESI DI DOTTORATO

ENGLISH ABSTRACT

Text Retrieval and Categorization Through a Weighted Word Pairs Approach

CANDIDATO: **LUCA GRECO**

TUTOR: **PROF. MASSIMO DE SANTO**

COORDINATORE: **PROF. ANGELO MARCELLI**

Anno Accademico 2011 – 2012

Text Retrieval and Categorization through a Weighted Word Pairs Approach

AUTHOR: LUCA GRECO

ADVISOR: PROF. MASSIMO DE SANTO

Abstract

The focus of this dissertation is the development and validation of a novel method for supervised text classification to be used effectively when small sized training sets are available. The proposed approach, which relies on a Weighted Word Pairs (WWP) structure, has been validated in two application fields: Query Expansion and Text Categorization.

By analyzing the state of the art for supervised text classification, it has been observed that existing methods show a drastic performance decrease when the number of training examples is reduced. This behaviour is essentialy due to the following reasons: the use, common to most existing systems, of the "Bag of Words" model where only the presence and occurrence of words in texts is considered, losing any information about the position; polysemy and ambiguity which are typical of natural language; the performance degradation affecting classification systems when the number of features is much greater than the available training samples.

Nevertheless, manual document classification is a boring, costly and slow process: it has been observed that only 100 documents can be hand-labeled in 90 minutes and this number may be not sufficient for achieving good accuracy in real contexts with a standard trained classifier. On the other hand, in Query Expansion problems (in the domain of interactive web search engines), where the user is asked to provide a relevance feedback to refine the search process, the number of selected documents is much less than the total number of indexed documents. Hence, there's a great interest in alternative classification methods which, using more complex structures than a simple list of words, show higher efficiency when learning from a few training documents.

The proposed approach is based on a hierarchical structure, called Weighted Word Pairs (WWP), that can be learned automatically from a corpus of documents and relies on two fundamental entities: *aggregate roots* i.e. the words probabilistically more implied from all others; *aggregates* which are words having a greater probabilistic correlation with aggregate roots.

WWP structure learning takes place through three main phases: the first phase is characterized by the use of probabilistic topic model and Latent Dirichlet Allocation to compute the probability distribution of words within documents: in particular, the output of LDA algorithm consists of two matrices that define the probabilistic relationship between words, topics and the documents. Under suitable assumptions, the probability of the occurrence of each word in the corpus, the conditional and joint probabilities between word pairs can be derived from these matrices. During the second phase, aggregate roots (whose number is selected by the user as an external parameter) are chosen as those words that maximize the conditional probability product between a given word and all others, in line with the definition given above. Once aggregate roots have been chosen, each of them is associated with some aggregates and the coefficient of relationship between aggregate roots and aggregates is calculated thanks to the joint probability between word pairs (previously computed). The number of links between aggregate roots and aggregates depends on another external parameter (Max Pairs) which affects proper thresholds allowing to filter weakly correlated pairs. The third phase is aimed at searching the optimal WWP structure, which has to provide a synthetic representation for the information contained in all the documents (not only into a subset of them).

The effectiveness of the WWP structure was initially assessed in Query Expansion problems, in the context of interactive search engines. In this scenario, the user, after getting from the system a first ranking of

documents in response to a specific query, is asked to select some relevant documents as a feedback, according to his information need. From those documents (relevance feedback), some key terms are extracted to expand the initial query and refine the search. In our case, a WWP structure is extracted from the relevance feedback and is appropriately translated into a query.
The experimental phase for this application context was conducted with the use of TREC-8 standard dataset, which consists of approximately 520 thousand pre-classified documents. A performance comparison between the baseline (results obtained with no expanded query), WWP structure and a query expansion method based on the Kullback Leibler divergence was carried out. Typical information retrieval measurement were computed: precision at various levels, mean average precision, binary preference, R-precision. The evaluation of these measurements was performed using a standard evaluation tool used for TREC conferences. The results obtained are very encouraging.

A further application field for validating WWP structure is documents categorization. In this case, a WWP structure combined with a standard Information Retrieval module is used to implement a document-ranking text classifier. Such a classifier is able to make a soft decision: it draws up a ranking of documents that requires the choice of an appropriate threshold (Categorization Status Value) in order to obtain a binary classification. In our case, this threshold was chosen by evaluating performance on a *validation set* in terms of micro-precision, micro-recall and micro-F1. The dataset Reuters-21578, consisting of about 21 thousand newspaper articles, has been used; in particular, evaluation was performed on the ModApte split (10 categories), which includes only manually classified documents. The experiment was carried out by selecting randomly the 1% of the training set available for each category and this selection was made 100 times so that the results were not biased by the specific subset. The performance, evaluated by calculating the F1 measure (harmonic mean of precision and recall), was compared with the Support Vector Machines, in the literature referred as the state of the art in the classification of such a dataset. The results show that when the training set is reduced to 1%, the performance of the classifier based on WWP are on average higher than those of SVM.