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

ABSTRACT ITALIANO

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

CANDIDATO: LUCA GRECO
TUTOR: PROF. MASSIMO DE SANTO

Dottorato di Ricerca in Ingegneria dell'Informazione XI Ciclo N.S.

Abstract

Il focus dell'attività di ricerca riguarda lo sviluppo e la validazione di una metodologia alternativa per la classificazione supervisionata di testi mediante impiego di training set di dimensioni ridotte (circa l'1% rispetto a quelli tipicamente impiegati). L'approccio proposto, che si basa su una struttura a coppie di parole pesate (Weighted Word Pairs), è stato validato su due contesti applicativi: Query Expansion e Text Categorization.

Da un'accurata analisi dello stato dell'arte in materia di classificazione supervisionata dei testi, si è evinto come le metodologie esistenti mostrino un evidente calo di prestazioni in presenza di una riduzione degli esempi (campioni del data set già classificati) utilizzati per l'addestramento. Tale calo è essenzialmente attribuibile alle seguenti cause: l'impiego, comune a gran parte dei sistemi esistenti, del modello "Bag of Words" dove si tiene conto della sola presenza ed occorrenza delle singole parole nei testi, perdendo qualsiasi informazione circa la posizione; polisemia ed ambiguità tipiche del linguaggio naturale; il peggioramento delle prestazioni che coinvolge i sistemi di classificazione quando il numero di caratteristiche (*features*) impiegate è molto maggiore degli esempi disponibili per l'addestramento del sistema.

Dal punto di vista delle applicazioni, ci si trova spesso di fronte a casi in cui, per la classificazione di un corpus di documenti, si ha a disposizione un insieme limitato di esempi: questo perché il processo di classificazione manuale dei documenti è oneroso e lento. D'altro canto in problemi di Query Expansion, nell'ambito dei motori di ricerca interattivi, dove l'utente è chiamato a fornire un feedback di rilevanza per raffinare il processo di ricerca, il numero di documenti selezionati è molto inferiore al totale dei documenti indicizzati dal motore. Da qui l'interesse verso strategie di classificazione che, usando strutture più complesse rispetto alla semplice lista di parole, mostrino un'efficienza maggiore quando la struttura è appresa da pochi documenti di training.

L'approccio proposto si basa su una struttura gerarchica (*Weighted Word Pairs*) che può essere appresa automaticamente da un corpus di documenti e che è costituita da due entità fondamentali: i termini *aggregatori* che sono le parole probabilisticamente più implicate da tutte le altre; i termini *aggregati* che sono le parole aventi maggiore correlazione probabilistica con i termini aggregatori.

L'apprendimento della struttura WWP avviene attraverso tre fasi principali: la prima fase è caratterizzata dall'impiego del topic model probabilistico e della Latent Dirichlet Allocation per il calcolo della distribuzione probabilistica delle parole all'interno dei documenti: in particolare, l'output dell'algoritmo LDA è costituito da due matrici che definiscono il legame probabilistico tra le parole, i topic e i documenti analizzati. Sotto opportune ipotesi è possibile derivare da tali matrici le probabilità associate al verificarsi delle singole parole all'interno del corpus e le probabilità condizionate e congiunte tra le coppie di parole; durante la seconda fase vengono scelti i termini *aggregatori* (il cui numero è selezionato dall'utente come parametro esterno) come quelle parole che massimizzano il prodotto delle probabilità condizionate al verificarsi di tutte le altre, coerentemente con la definizione fornita in precedenza. Una volta scelti i termini aggregatori, a ciascuno di essi sono associati dei termini aggregati e il coefficiente di relazione tra termini aggregatori ed aggregati è calcolato sulla base della probabilità congiunta. Il numero di legami tra aggregatori e tra aggregatori/aggregati dipende da un parametro esterno (Max Pairs) che va ad influire su opportune soglie che filtrano le coppie debolmente correlate. La terza fase ha come obiettivo la ricerca della struttura WWP ottima, che tenga conto dell'informazione presente in tutti i documenti del corpus e che non sia maggiormente caratterizzata da un sottoinsieme di essi.

L'efficacia della struttura WWP è stata dapprima valutata in problemi di Query Expansion nell'ambito dei motori di ricerca interattivi. In questo scenario l'utente, dopo aver ottenuto dal sistema un primo ranking di documenti in risposta ad una sua query iniziale, è chiamato a selezionare alcuni documenti da lui giudicati

rilevanti che andranno a costituire il *relevance feedback* da cui estrarre opportunamente nuovi termini per espandere la query iniziale e raffinare la ricerca. Nel caso specifico, la struttura WWP appresa dal relevance feedback viene opportunamente tradotta in una query mediante un linguaggio di interrogazione proprio del modulo di Information Retrieval utilizzato.
La sperimentazione in questo contesto applicativo è stata condotta mediante l'utilizzo del dataset standard TREC-8, costituito da circa 520 mila documenti pre-classificati. E' stato effettuato un confronto di performance tra la baseline (risultati ottenuti da query priva di espansione), la struttura WWP ed un metodo di espansione basato sulla Divergenza di Kullback Leibler, indicato in letteratura come il metodo di estrazione delle feature più performante nei problemi di query expansion; le misurazioni effettuate sono tipiche dell'information retrieval: precisione a vari livelli, *mean average precision*, *binary preference*, R-precision. La valutazione di tali quantità è stata effettuata utilizzando un apposito tool messo a disposizione per la conferenza TREC. I risultati ottenuti sono molto incoraggianti.
Un ulteriore campo applicativo in cui la struttura è stata validata è quello della categorizzazione dei documenti. In questo caso, la struttura WWP abbinata ad un modulo di Information Retrieval è utilizzata per implementare un *document-ranking text classifier*. Un classificatore di questo tipo realizza una *soft decision* ovvero non fornisce in ouput l'appartenenza di un documento ad una determinata classe ma redige un ranking di documenti che richiede la scelta di un opportuna soglia (*Categorization Status Value threshold)* per consentire la classificazione vera e propria. Tale soglia è stata scelta valutando le performance del classificatore in termini di micro-precision, micro-recall e micro-F1 rispetto al dataset utilizzato. Quest'ultimo, noto in letteratura come Reuters-21578, è costituito da circa 21 mila articoli di giornale; il sottoinsieme utilizzato, noto come ModApte split, include i soli documenti classificati manualmente da umani (10 categorie). La sperimentazione è stata condotta selezionando l'1% in maniera random del training set di ciascuna categoria e tale selezione è stata effettuata 100 volte in modo che i risultati non fossero polarizzati dallo specifico sottoinsieme. Le performance, valutate mediante calcolo della misura F1 (media armonica di precisione e richiamo), sono state confrontate con le Support Vector Machines, in letteratura indicate come stato dell'arte nella classificazione del dataset impiegato. I risultati ottenuti mostrano che quando il training set è ridotto al 1%, le performance del classificatore basato su WWP sono mediamente superiori a quelle delle SVM.
I risultati ottenuti dall'impiego della struttura WWP nei campi di Text Retrieval e Text Mining sono molto interessanti e stanno ottenendo buon riscontro da parte della comunità scientifica. Dal punto di vista delle prospettive future, essendo attualmente la struttura appresa dai soli esempi positivi, potrebbe essere interessante valutare l'incremento di performance ottenuto impiegando 2 strutture WWP apprese rispettivamente da esempi positivi e negativi. Naturalmente, trattandosi di un classificatore *soft decision*, diventa cruciale stabilire una corretta politica di combinazione tra i ranking ottenuti dall'impiego del WWP "positivo" e quello negativo e la scelta della soglia CSV.
Un altro interessante spunto futuro riguarda la costruzione di ontologie complete da strutture WWP, che richiederebbe l'identificazione delle tipologie di relazioni esistenti tra i termini mediante ausilio di conoscenza esogena (WordNet, etc...).