Strategie di classificazione per servizi di search della Pubblica Amministrazione

Marco Bianchi[1], Mauro Draoli[2] Giorgio Gambosi[1], Alessandro Ligi[2], and Marco Serrago

[1] University of Rome "Tor Vergata",
Via della Ricerca Scientifica 1, 00133 Rome, Italy
bianchi@mat.uniroma2.it
gambosi@mat.uniroma2.it
[2] DigitPA, Viale Marx 43 - 00137 Rome, Italy
draoli@digitpa.gov.it
alessandro.ligi@digitpa.gov.it

Sommario In questo lavoro si introducono le attività sperimentali finalizzate alla realizzazione di un servizio per la ricerca della modulistica pubblicata dalle Pubbliche Amministrazioni (PA) italiane sui propri siti istituzionali e condotte nell'ambito del progetto pubblico "Italia.gov.it - il motore della PA digitale". In tale contesto la necessità di creare e aggiornare una collezione composta da soli moduli rende necessaria l'introduzione di classificatori automatici che siano in grado di supportare il filtering della grande mole di documenti che vengono recuperati a valle dell'attività di crawling. Il caso presentato è interessante perchè mostra quanto la scelta del classificatore da adottare possa essere influenzata dai vincoli economici e organizzativi tipicamente posti dalle Pubbliche Amministrazioni.

Keywords: vertical search engine, classification, active learning.

Oggigiorno la realizzazione di servizi di *search* verticali per il dominio della Pubblica Amministrazione (PA) è un'attività il cui valore scientifico, economico e sociale viene riconosciuto a livello internazionale. Il motore di ricerca USA.GOV[3] è forse il principale esempio di applicazioni in esercizio con l'obiettivo di supportare i cittadini e le aziende nella ricerca di informazioni e documenti pubblicati sul Web dalla PA. Anche in Italia è stato avviato un progetto pubblico finalizzato alla realizzazione di un motore di ricerca della PA, denominato Italia.gov.it[4]. Nell'ambito di questo progetto ogni funzionalità di *search* definisce, di fatto, un *task* a sé stante che spesso richiede la sperimentazione di soluzioni innovative.

In questo lavoro si illustrano alcune problematiche che si stanno affrontando durante le attività sperimentali finalizzate alla realizzazione di un servizio per la ricerca della modulistica pubblicata dalle PA sui propri siti istituzionali. Tale servizio, denominato *moduli-on-line*, rappresenta un esempio significativo

[3] http://search.usa.gov/
[4] http://www.italia.gov.it

delle funzionalità di *search* erogate da Italia.gov.it e di come esse possono essere implementate.

Uno degli aspetti innovativi di Italia.gov.it risiede nella presenza di una base di conoscenza da cui si attingono tutte le informazioni che vengono indicizzate per la realizzazione dei singoli servizi di *search*. Tale base di conoscenza è caratterizzata da una modalità di aggiornamento automatico eseguita per mezzo di strumenti di Information Retrieval e Text Mining allo stato dell'arte. Nello specifico, il servizio moduli-on-line indicizza tutti i documenti scoperti sul Web per mezzo di una continua attività di crawling e marcati come *moduli* da classificatori binari precedentemente addestrati. Il lavoro svolto dai classificatori è pertanto determinante per ottenere una buona qualità degli indici di ricerca: se i classificatori svolgono il proprio lavoro con precisione i risultati presentati agli utenti saranno composti, per lo più, da modulistica, viceversa saranno "inquinati" da errori di classificazione (falsi positivi) che potrebbero minare alla base la fiducia sul funzionamento del sistema. Poiché la qualità del servizio che si intende realizzare è così fortemente influenzata dal funzionamento dei classificatori, è stata avviata un'attività finalizzata alla creazione di un benchmark utile per l'addestramento, la misurazione e la scelta del miglior tipo di classificatore per il problema in oggetto.

In questo lavoro si descrive la strategia che si sta studiando per la gestione del servizio moduli-on-line, conciliando esigenze di natura sia economica che tecnica. Per quanto riguarda le esigenze economiche, l'architettura del sistema Italia.gov.it, descritta in [1], prevede il possibile intervento di esperti di dominio (oracoli, dal punto di vista del processo di classificazione) che da un lato eseguono un monitoraggio continuo sulla precisione del sistema di classificazione, dall'altro risolvono i casi di incertezza degli strumenti di classificazione contribuendo, di fatto, a un arricchimento del training set [5]. In questo scenario, si richiede che, in fase di produzione, gli oracoli siano messi nella condizione di classificare quanti più moduli possibile, al fine di massimizzare il numero di documenti indicizzati. È evidente, infatti, come la classificazione di un non-modulo non sia immediatamente riutilizzabile in fase di produzione. Dal punto di vista tecnico-scientifico, invece, il training set deve essere rappresentativo del dominio di classificazione e il suo aggiornamento deve essere finalizzato al miglioramento delle prestazioni dei classificatori. Pertanto, anche gli esempi veri negativi e falsi positivi possono essere utili per migliorare il sistema di classificazione. È obiettivo della sperimentazione in corso verificare la compatibilità tra le due esigenze.

L'attività di sperimentazione è condotta concentrando l'attenzione su classificatori di tipo Support Vector Machine (SVM) [2], Naive Bayes (NB) [4], Logistic Regression (LR) [3] e Dynamic Language Model (DLM) [6].

La prima versione del benchmark di moduli-on-line, è composta da 8475 documenti recuperati a valle di un'attività di crawling eseguita nel mese di marzo 2011 su 12 siti istituzionali di PA centrali indicati da esperti di dominio. Il crawler è stato configurato in modo da scaricare solamente documenti con estensioni .pdf, .doc, .docx, .rtf, .xls e .xlsx in quanto si pensa possano essere i principali formati utilizzati dalle PA per la pubblicazione della modulistica. L'intero insieme dei

file è stato successivamente classificato a mano da esperti di dominio che hanno individuato 793 documenti appartenenti alla categoria dei moduli e 7461 a quella dei non-moduli. Per la fase di valutazione si è considerato come *modulo* ogni documento testuale realizzato per scopi amministrativi e burocratici comprensivo di una serie di campi compilabili da un generico utente. Nella classe complementare sono invece stato inseriti tutti gli altri documenti. Esempi tipici di documenti classificati come non-moduli sono le Determinazioni Dirigenziali, le Disposizioni Direttoriali, le Leggi, i Decreti Legge, gli Avvisi Pubblici. Si evidenzia la presenza di casi che potrebbero essere considerati di ambiguità. Esistono infatti documenti che sono caratterizzati dalla presenza di una prima parte documentale, e una seconda parte compilabile. Un esempio di questa tipologia di documenti è un Bando di Concorso che è tipicamente composto da un certo numero di articoli che regolamentano la procedura concorsuale e da alcuni modelli di modulo in appendice. Su indicazione degli esperti di dominio, i casi di incertezza sono stati aggiunti alla categoria dei moduli.

Un primo training set delle dimensioni di 547 documenti (composto da 325 documenti e 222 moduli) è stato costruito da esperti di dominio. La Tabella 1 riporta le prestazioni dei classificatori presi in esame sul test-set composto dai rimanenti 7707 documenti.

Classif.	Precision	FP-rate	Recall	Accuracy	F-measure
DLM	0.88	0.01	0.82	0.97	0.84
LR	0.79	0.02	0.58	0.92	0.67
SVM	0.66	0.04	0.70	0.94	0.68
NB	0.51	0.05	0.71	0.93	0.60

Tabella 1. Tabella che riassume le prestazioni dei classificatori analizzati. La stabilità dei risultati è stata verificata mediante cross-validation.

A una prima analisi, il classificatore che sembra fornire le migliori prestazioni è il DLM con una precisione dell'88% e una recall dell'82%. Purtroppo però, per requisiti di progetto, la precisione di classificazione deve superare il 95% anche a costo di coinvolgere gli oracoli nel processo di classificazione. Di conseguenza il problema diventa individuare quell'insieme di documenti da sottomettere agli oracoli al fine di superare il 95% di precisione, non penalizzando eccessivamente la recall e minimizzare il lavoro manuale svolto dagli oracoli.

Per affrontare questo nuovo problema si è pensato di osservare i valori di probabilità che i singoli classificatori assegnano ai documenti al fine di fornire indicazione sul grado di confidenza con il quale determinano l'appartenenza a una classe. La Tabella 2 mostra alcuni dettagli sul comportamento dei classificatori DLM e SVM. Analizzando gli insiemi dei documenti classificati come moduli si osserva come il classificatore DLM, assegni la classe di appartenenza con probabilità maggiore del 95% in ben 7680 casi; diversamente SVM ha un comportamento che potremmo definire più cauto, assegnando solo in 3786 casi una probabilità maggiore del 95%. Fissato un intervallo $(x-5, x]$ consideriamo il seguente processo semi-automatico di classificazione:

1. Tutti i documenti identificati come moduli dal classificatore con probabilità maggiore di $x\%$ sono accettati come tali;

2. I documenti classificati come moduli con probabilità compresa tra 50% e x% sono sottoposti alla valutazione manuale assumendo che tale valutazione abbia errore nullo.

Denotiamo come $P_{inc}(x)$ la precisione derivante da questo processo semi-automatico di classificazione. Le stesse considerazioni possono essere effettuate riguardo la recall indicata con $R_{inc}(x)$. La Tabella 2 mostra come considerando una strategia di classificazione semi-automatica, il classificatore SVN sia da considerarsi preferibile in quanto mette il decisore finale nella condizione di poter "acquistare" la precisione voluta pagando il costo di valutazione manuale delle classificazioni incerte (es. 466 valutazioni per raggiungere il 95.7% di precisione). Si evidenzia che se si è disposti a perdere in recall, la precisione può addirittura aumentare riducendo il costo di valutazione manuale. Ciò è possibile passando agli oracoli un pacchetto di documenti selezionato negli intervalli di confidenza con probabilità più alta. È obiettivo di questa sperimentazione verificare che questa modalità di scelta dei documenti da passare agli oracoli non penalizzi le prestazioni dei classificatori successivamente alle attività di ri-addestramento. A tal fine si è deciso di incrementare sensibilmente la dimensione del benchmark, che sarà ampliato fino a raggiungere circa 30.000 documenti classificati a mano.

	(50,55]	(55,60]	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]	(85,90]	(90,95]	(95,100]
DLM										
TP+FP	1	2	0	0	2	0	1	2	3	841
TN+FN	0	3	2	2	0	1	2	3	5	6839
P_{inc}	0.876	0.877	0.877	0.877	0.878	0.878	0.878	0.878	0.879	1
R_{inc}	0.816	0.819	0.819	0.819	0.821	0.821	0.822	0.825	0.827	1
SVM										
TP+FP	98	68	60	47	46	50	31	36	30	284
TN+FN	99	150	199	237	316	415	584	620	837	3502
P_{inc}	0.706	0.749	0.785	0.815	0.846	0.880	0.914	0.940	0.957	1
R_{inc}	0.782	0.846	0.881	0.902	0.926	0.954	0.958	0.969	0.972	1

Tabella 2. Dettaglio dei comportamenti dei classificatori DLM e SVM.

Riferimenti bibliografici

1. Bianchi, M., Draoli, M., and Gambosi, G. An innovative approach to the development of e-government search services. In *EGOVIS* (2011), K. N. Andersen, E. Francesconi, Å. Grönlund, and T. M. van Engers, Eds., vol. 6866 of *Lecture Notes in Computer Science*, Springer, pp. 41–55.
2. Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20* (1995), 273–297. 10.1007/BF00994018.
3. Hosmer, D. W., and Lemeshow, S. *Applied logistic regression (Wiley Series in probability and statistics)*, 2 ed. Wiley-Interscience Publication, 2000.
4. John, G., and Langley, P. Estimating continuous distributions in bayesian classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (1995), Morgan Kaufmann, pp. 338–345.
5. Settles, B. Active learning literature survey. Tech. rep., 2010.
6. Zhai, C. *Statistical Language Models for Information Retrieval*. Now Publishers Inc., Hanover, MA, USA, 2008.