v.3

ECoRe: Enterprise Content Recommendation

Daniele Turato, Alessandro Berti, Marco Pegoraro • 5/10/2017

1. Sommario

2. Obiettivo del progetto

L'obiettivo finale del progetto è la realizzazione di un servizio proattivo in grado di suggerire all'utente che accede a contenuti aziendali (tramite vari punti di accesso: email, documentale[1], ecc.), altri contenuti di interesse che potrebbero essere utili nello svolgimento del proprio lavoro. Tale utilità sarà stabilita sulla base del comportamento dell'utente stesso.

2.1. Esempio esplicativo

Un sales riceve un messaggio di posta dal cliente CPharma Spa, riguardo alla richiesta di preventivo per il nuovo contratto di manutenzione sul software documentale. All'apertura del messaggio, il sistema provvede a pubblicare nella barra a piè di pagina i seguenti suggerimenti:

- Notifica della chiusura di 2 ticket nell'ultima settimana per il cliente CPharma
- Notifica della nuova opportunità sul cliente KLogistics Spa riguardo ai contratti di manutenzione sullo stesso software
- 2 documenti di interesse:
    - Guida ai contratti di manutenzione
    - Progetto gestione protocollo CPharma
- Notizie dal web: CPharma lancia un nuovo antibiotico rivoluzionario

3. Dominio applicativo

3.1. Enterprise search e Sistemi di raccomandazione

Con il termine enterprise search si intende l'applicazione delle conoscenze e dei metodi del reperimento dell'informazione nel contesto della ricerca di informazioni all'interno di una organizzazione aziendale.
La caratteristica peculiare è che la base documentale è costituita non da un'unica sorgente, ma da diverse sorgenti, quali ad esempio il web (spesso specificando delle repository scelte), l'intranet interna, e ogni altra fonte di testi digitale dell'organizzazione (messaggi di posta elettronica, database, log di eventi, ecc). Di conseguenza, la base documentale può essere costituita da dati presenti sia in formati strutturati che non strutturati. Inoltre l'accesso ai documenti è dipendente da ruoli e autorizzazioni di chi utilizza il sistema all'interno dell'azienda [1].

Tra i motori open-source di Enterprise Search più conosciuti possiamo citare:

- Apache SolR[2]
- Elasticsearch[3]

I sistemi di raccomandazione sono strumenti software in grado di suggerire contenuti utili all'utente [2]. Nell'accezione più generica del termine, la caratteristica distintiva rispetto ai motori di ricerca è che non viene richiesto all'utente l'accesso ad una specifica interfaccia di ricerca per esprimere una query esplicita. Le differenze tuttavia si esplicano anche nella tecniche utilizzate per erogare le rispettive funzionalità. All'interno dei sistemi di raccomandazione, la classificazione principale è basata sull'approccio utilizzato. In particolare si distinguono:

- Collaborative Filtering: si cerca di privilegiare nella raccomandazione i contenuti che sono risultati in passato interessanti per utenti con gusti simili rispetto a quello corrente

---

[1] Un sistema di gestione documentale (DMS, Document Management System) è un sistema informativo progettato per tracciare, gestire e archiviare documenti e ridurre la carta. Viene spesso considerato come un componente dei sistemi di gestione dei contenuti aziendali (detti ECM, Enterprise Content Management).
[2] http://lucene.apache.org/solr/
[3] https://www.elastic.co/

- Content-based Filtering: si cerca di privilegiare nella raccomandazione i contenuti simili a quelli che l'utente ha ritenuto interessanti nel passato.

Come punto di partenza, abbiamo pensato di utilizzare un motore di ricerca come SolR o ElasticSearch perché mettono a disposizione:

- Funzionalità di importazione dati da molteplici sistemi informativi tramite handler specifici o servizi rest richiamabili da moduli esterni
- Funzionalità di base necessarie al preprocessing del documento per ridurne la dimensionalità e efficientarne l'indicizzazione e lo storage
- Indicizzazione dei documenti (tramite Lucene[4])
- Funzionalità che permettono di erogare una prima semplice forma di raccomandazione (tramite MoreLikeThese o similari)

Inoltre, la loro architettura è studiata per essere facilmente estensibile, e la proponente li utilizza già in altri progetti aziendali.

La parte centrale del progetto verte quindi sul come utilizzare al meglio questi strumenti e stabilire quali moduli aggiuntivi andranno implementati per soddisfare i requisiti. Potremmo più propriamente definire l'obiettivo del progetto come un **sistema di enterprise search con funzionalità di raccomandazione**.

Nella realizzazione del progetto potrà essere utile anche l'esperienza maturata con precedenti lavori in collaborazione con l'università di Padova [3,4].

3.2. Tipologie di contenuti

Le tipologie di contenuti su cui desideriamo implementare un sistema di raccomandazione sono:

- Documenti di interesse aziendale: tra questi possiamo citare fatture, contratti, procedure aziendali, materiale marketing, etc. Da questi documenti, contenuti nel loro formato originario all'interno del sistema documentale aziendale, sarà estratto il solo testo, in modo da poterli inserire nel sistema (esistono già molti strumenti proprietari della proponente e non per implementare questa parte).
- Eventi: un evento è la registrazione di attività in un sistema informativo aziendale. Di particolare interesse sono gli eventi che fanno riferimento a un processo aziendale[5] (ad esempio: gestione dei ticket, gestione delle opportunità commerciali). Un evento appartiene a una specifica istanza del processo (ad esempio, nel contesto di gestione dei ticket di assistenza, ogni singolo ticket avrà un evento di "presa in carico" o di "assegnazione priorità").
- Contenuti web: tramite l'utilizzo di tecniche di web crawling, è possibile reperire il contenuto di siti web specifici. Possiamo citare: siti web di interesse normativo, indagini di settore, news specialistiche, siti web dei clienti.

4. Sistema di raccomandazione per contenuti aziendali

Lo scopo finale del progetto è la creazione di un sistema che sia in grado di suggerire all'utente contenuti utili a semplificare e rendere più efficace il suo lavoro. Esempi di utente di riferimento sono le figure disales, business consultant, marketing. Il caso d'uso non è definito a priori e dovrà essere stabilito sulla

[4] (https://lucene.apache.org) API gratuita ed open source per il reperimento di informazioni, concepita inizialmente per realizzare applicazioni che necessitano di funzionalità di indicizzazione e ricerca full text, è ora molto nota ed usata per la realizzazione di motori di ricerca sia sul World Wide Web che sulle Intranet private.

[5] "Un processo è un insieme di attività strutturate e misurate, progettato per produrre uno specifico output per un mercato o un cliente particolare" [5]

base di diversi criteri (opportunità commerciali, fattibilità nei tempi del progetto, applicabilità in azienda ecc.). Alcune ipotesi sono:

- Integrazione in un sistema documentale: nel momento in cui una scheda documento viene visualizzata, si suggeriscono contenuti correlati al documento principale e/o agli allegati della scheda documento.
- Integrazione in un sistema di Customer relationship management (CRM)[6]
- Realizzazione di un plugin per un client di posta elettronica: nel momento della visualizzazione di un messaggio di posta, si vanno a suggerire contenuti correlati all'oggetto, al testo del messaggio e/o al contenuto degli allegati.
- Realizzazione di un'applicazione Android: nel momento in cui arriva una notifica relativa ad un messaggio di posta, si vanno a proporre i contenuti correlati.

L'interfaccia di visualizzazione dei contenuti raccomandati dovrà permettere di visualizzarli tramite l'utilizzo di visualizzatori forniti dagli stessi sistemi target, oppure, nel caso questo non sia possibile, tramite un visualizzatore delle sole informazioni indicizzate all'interno del sistema di raccomandazione. Si ipotizza anche che l'interfaccia suggerisca ulteriori contenuti affini ai suggerimenti stessi man mano che questi vengono esplorati.
Al fine di poter affinare la qualità delle raccomandazioni, si dovrà tener conto della storia dell'utente, ossia dei contenuti consultati e dei suggerimenti che sono risultati interessanti per l'utente.
Per poter indicizzare i contenuti appartenenti alle tipologie di interesse, dovranno essere implementate metodologie di estrazione necessarie a caricare tali contenuti nel sistema. Tali metodologie potranno avvalersi degli strumenti forniti dagli eventuali motori di enterprise search utilizzati, oppure sfruttare le librerie già preparate dalla proponente in altri progetti.
In vista dei test del sistema, al fornitore verrà dato accesso a macchine virtuali da cui sarà possibile accedere a contenuti reali appartenenti alle tipologie di interesse da utilizzare, al momento opportuno e con metodiche che rispettino gli standard di sicurezza. Sia i contenuti, sia l'indice del motore, dovranno rimanere (per ragioni di privacy) all'interno dei sistemi aziendali della proponente.

5. Specifiche di progetto e riferimenti

5.1. Scelte tecnologiche

Ipotizziamo che il motore di raccomandazione possa essere esposto come una serie di servizi web accessibili dal contesto in cui la raccomandazione sarà abilitata. Quindi, l'accessibilità degli stessi sarà dipendente dal caso d'uso concordato.

Come già segnalato, si raccomanda l'utilizzo, come base di partenza per la costruzione del sistema, di uno dei seguenti progetti open source di Enterprise Search:

- Apache SolR
- Elasticsearch

Questi dovranno essere integrati con l'implementazione dei necessari moduli che eseguono la raccomandazione, dove potrebbe tornare utile la libreria di apprendimento automatico Apache Mahout[7], che contiene anche una parte dedicata appositamente alla raccomandazione.

Nei casi in cui sia necessario l'accesso esterno ai servizi web (ad esempio, nel caso d'uso mobile), dovranno essere previste anche adeguate misure di sicurezza, quali:

- Esposizione dei servizi in modalità HTTPS

[6] Le applicazioni CRM servono a tenersi in contatto con la clientela, a inserire le loro informazioni nel database e a fornire loro modalità per interagire in modo che tali interazioni possano essere registrate e analizzate.
[7] https://mahout.apache.org/

- Interfacciamento al sistema di Identity and Access Management Keycloak[8] per l'autenticazione degli utenti.

Per l'importazione di dati provenienti da fonti web, si raccomanda di usare Apache Nutch[9], un motore di ricerca open source altamente estensibile che la proponente ha già utilizzato con successo in passato.

Nella scelta delle componenti software e librerie da utilizzare nella realizzazione del progetto, dovrà essere posta massima attenzione alla compatibilità con l'ecosistema di prodotti della proponente. Questo vincolo non pone limiti in termini di utilizzo di componenti con licenza libera, a condizione che esse non limitino o impediscano l'utilizzo commerciale del software risultante dal progetto.
Alcune librerie di supporto già sviluppate dalla proponente potrebbero essere utilizzate nel contesto del progetto. Tuttavia, in tal caso, il codice deve rispettare le licenze stabilite dall'azienda stessa. In molti casi si potrà optare per l'utilizzo di analoghe librerie open, a condizione che il codice sia strutturato in modo da permettere una loro agevole sostituzione con controparti proprietarie, dove ciò sia ritenuto migliorativo. Un esempio in proposito è la parte di estrazione del testo, dove si potrà usare la libreria open source Apache Tika[10], ma dove la proponente ha già sviluppato strumenti alternativi più avanzati.

5.2. Obiettivi desiderabili

- Realizzazione del motore di raccomandazione (backend)
  - Esposizione servizi di raccomandazione
  - Configurazione o implementazione dei necessari connettori:
    - Documenti (solo configurazione)
    - Eventi e opportunità helpdesk (da Microsoft Dynamics tramite strato OData[11])
  - Gestione della visibilità dei contenuti a seconda del ruolo dell'utente destinatario
- Individuazione di un caso d'uso di fruizione delle raccomandazioni; alcuni esempi:
  - Dal visualizzatore del documento nel software di gestione documentale
  - Durante la lettura di un messaggio dal client di posta elettronica (es. Outlook, Thunderbird)
  - Durante la navigazione da browser
  - Tutti i precedenti ma da smartphone
- Realizzazione di un'applicazione che permetta di fruire dei servizi di raccomandazione
- Documentazione dell'applicazione (in base al caso d'uso scelto)
  - analisi dei requisiti;
  - motivazione delle scelte tecnologiche;
  - descrizione tecnica.

Una nota a parte merita la gestione della documentazione. Esclusa la documentazione prevista dalle regole del capitolato, , la proponente vincola all'utilizzo di Evernote[12], come strumento per ospitare la base di conoscenza tecnica relativa al progetto. Con questo si intende che il fornitore dovrà documentare in Evernote ogni contenuto concordato come significativo intorno al progetto, in particolare:

- Scelte tecniche e implementative

---

[8] http://www.keycloak.org/
[9] http://nutch.apache.org/
[10] (https://tika.apache.org/) è un toolkit (disponibile sia come libreria che come applicazione a riga di comando) che rileva ed estrae metadati e testo da oltre mille tipi di file diversi
[11] (Open Data Protocol) protocollo aperto che permette la creazione e il consumo di API RESTful interoperabili e interrogabili in maniera semplice e standard, utilizzato per fornire accesso ai dati contenuti in un sistema informativo
[12] (https://evernote.com/intl/it/) è un'applicazione cross-platform progettata per la scrittura, l'organizzazione e l'archiviazione di note. Evernote supporta la maggior parte delle piattaforme di sistema operative popolari.

- Output degli incontri
- Difficoltà incontrate e soluzioni adottate
- Problemi aperti e possibili soluzioni
- Guida all'installazione ed esecuzione sia in ottica di deployment, che di ricostruzione dell'ambiente di sviluppo.

Questa fornitura ha pari importanza a tutti gli altri obiettivi, ed è vitale ai fini dell'auspicabile uso futuro dei risultati del progetto.

5.3. Obiettivi aggiuntivi

- Fruizione delle raccomandazione in ambiente mobile
- Sicurezza dei dati memorizzati all'interno del sistema di raccomandazione
- Autenticazione dei servizi tramite Keycloak
- Valutazione di scalabilità (test di carico con numero realistico di utenti)

6. Note sul capitolato d'appalto

6.1. Aspettative della proponente

L'aspettativa minima della proponente è l'ottenimento di un prototipo software che:

- implementi i connettori Enterprise Search rispetto ai sistemi concordati
- esponga i servizi web di raccomandazione
- Implementi l'interfaccia rispetto al caso d'uso concordato.

Ci si aspetta che il prototipo software sia pronto per test dimostrativi interni al reparto di ricerca e sviluppo, per un numero ridotto di utenti contemporanei e senza la necessità di esporre i servizi di raccomandazione all'esterno della rete di Siav.

L'aspettativa ottima è l'ottenimento di un prototipo software avanzato pronto a test con un numero elevato di utenti contemporanei e in contesti reali.

Il fornitore dovrà produrre, e rilasciare all'azienda, documentazione su:

- Scelte implementative e progettuali effettuate e relative motivazioni
- Problemi aperti e eventuali soluzioni proposte da esplorare

6.2. Contatti

Il supporto della proponente verrà fornito attraverso i seguenti canali:

- Telegram (principale): sarà creato un gruppo apposito in modo da agevolare uno scambio più agile di informazioni dove questo sia necessario; questa sarà lo strumento di comunicazione preferenziale in quanto già adottato dal gruppo di ricerca e sviluppo di Siav.
- Telefono: 0498175246
- E-mail: daniele.turato@siav.it

Una persona del reparto di ricerca e sviluppo sarà incaricata di supervisionare l'avanzamento del progetto, essendo sempre disponibile, salvo impegni inderogabili, a interfacciarsi con il gruppo per eventuali evenienze.

6.3. Ambiente di sviluppo e strumenti

Nella parte di avvio del progetto saranno stabilite le modalità e gli strumenti più opportuni per agevolare il lavoro. Sulle base delle esperienze precedenti, è auspicabile che il fornitore si organizzi con un proprio repository esterno (Bitbucket/Github) e realizzi il lavoro in modo indipendente su postazioni proprie, così da svincolare il fornitore dal recarsi presso la sede della proponente quando non strettamente necessario. Durante lo sviluppo, il gruppo avrà necessariamente bisogno di simulare determinate funzionalità dei

moduli sviluppati contestualmente alla loro realizzazione. In tal caso, sempre all'inizio del progetto, si dovrà valutare di produrre un dataset di documenti prodotti sinteticamente o tramite anonimizzazione di documenti reali, allo scopo di permettere tale operazione di test garantendo allo stesso tempo la tutela dei dati dell'azienda. L'alternativa a questo tipo di operazione è quella di effettuare il prima possibile il deploy del motore di raccomandazione presso una macchina virtuale fornita dalla proponente, e effettuare l'estrazione delle informazioni dai diversi sistemi informativi target. Tale operazione però assume che si sia già valutata la sicurezza dell'architettura sempre con l'obiettivo della tutela dei dati dell'azienda, valutando i rischi in caso di violazione della macchina stessa. I dati della proponente non devono assolutamente uscire dalla macchina virtuale dedicata ad ospitare il motore di raccomandazione.

7. Note sulla proponente

Siav è una delle più importanti realtà italiane di sviluppo software e di servizi informatici specializzata nella dematerializzazione e nella gestione documentale e nei processi digitali.
Siav offre soluzioni in ambito di gestione documentale e dei processi, conservazione digitale per privati e PA, servizi in cloud e SaaS, dematerializzazione dei documenti cartacei, gestione dei processi di e-Government, sportello online del cittadino, archiviazione dei documenti digitali, gestione enterprise della PEC, File Sync & Sharing. La divisione specializzata in Document Management Outsourcing completa la gamma delle soluzioni offerte con servizi quali la Fatturazione alla PA e la Conservazione Digitale.
Siav è la principale azienda Italiana per l'Enterprise Content Management ed è fra le poche Accreditate dall'Agenzia per l'Italia Digitale.

8. Riferimenti bibliografici

1. Hawking D. Challenges in enterprise search. In: Proceedings of the 15th Australasian database conference - Volume 27. Australian Computer Society, Inc.; 2004. p. 15–24.

2. Ricci F, Rokach L, Shapira B. Introduction to Recommender Systems Handbook. In: Recommender Systems Handbook. 2010. p. 1–35.

3. Zilio D. Progettazione e realizzazione di un sistema di enterprise search. 2016.

4. Daniel Zilio Maristella Agosti. An Approach to the Design and Evaluation of an Enterprise Search Application. In: Proceedings of the 8th Italian Information Retrieval Workshop, Lugano, Switzerland, June 05-07, 2017. p. 17–23.

5. Davenport TH. Process Innovation: Reengineering Work Through Information Technology. Harvard Business Press; 2013. 352 p.