Genova, Feb 2013
Rev: 2.0
Confidential matter

CAIMANS

Schema fattuale del progetto preliminare

White sheet

Razionale progetto

Si intende realizzare una piattaforma proprietaria di classificazione automatica (**PPCA**) che sia in grado di effettuare la categorizzazione automatica di repertori di contenuti editoriali

I repertori sono attinti da volumi predisponibili di banche dati strutturate secondo un banking Entità-Relazione noti (**ERDB**). I contenuti afferenti alla piattaforma sono consumati in modo sistematico ma ordinalmente destrutturato in quanto la processazione del *dump* di un campo BLOB può reiterare sullo stesso record anche concorrenzialmente.

Scopo della PPCA è quello di originare una base di conoscenza associata (**AKB** - Associated Knowledge Base) per la categorizzazione di qualunque fonte testuale (TC - Text Categorisation). La AKB si definisce come "associata" intendendo con ciò il fatto che viene accoppiata ad un Document Data Base Server (**D-DBS**) che rappresenta il paradigma dei Training Sets di conoscenza. Pertanto, la AKB fornirà la struttura di meta-conoscenza D-DBS specifica fruibile per erogare tipologie diverse di servizi di Information Retrieval (**IRS**) legati ai contenuti complessivamente e concorrenzialmente classificati.

Al tempo della ideazione progettuale nessuna indicazione specifica è fornita sulla tipologia di servizi che si intende sviluppare sulla AKB, quindi tutte le possibilità di dossieraggio, statistiche, semplici inferenze e/o ricerche tematiche sono considerate. Allo stesso tempo nessuna forma evoluta di Full-Text Search (**FTS**) è legata all'uso della AKB e questa tecnica di ricerca del dato/informazione viene semplicemente pensata come ovvia in relazione allo standard del DB Server Engine che verrà adottato come stanrdad per il *repository* storico del D-DBS.Tuttavia, nel medio futuro, la FTS potrà essere combinata con i risultati della IRS o, addirittura integrata con i risultati di inferenze sulla AKB

Paternità e conio logotipi

Il nome del progetto CAIMANS è coniato dal progettista della piattaforma secondo un acronimo che successivamente alla fase di "prototipo operativo", potrà essere depositato e utilizzato per finalità di business congiuntamente con la committente.

L'acronimo : il nome del progetto "**CAIMANS**" è l'acronimo di *Classificazione Automatica di Istanze Multiple di Archivi Non Strutturati*.

Lo stesso nome è pertanto indicativo delle caratteristiche funzionali del sistema che nasce con lo scopo di creare una AKB e formalizza funzionalmente la nomenclatura di una struttura classificata proprietaria a partire da un aspecifico numero di contenuti testuali

La consulente svolgerà anche attività di editoria grigia e di settore pubblicando articoli originali per i quali i loghi di prodotto depositati ed eventualmente di società commerciali potranno essere liberamente citati e riprodotti.

La committente che avesse limitazioni e/o pregiudiziali nella capacità editoriale divulgativa, dovrà concordare una policy di rilascio e sfruttamento editoriale prima della conclusione della fase di beta-test.

La consulente, in quanto autore, è disponibili a considerare prima della fase sopracitata, qualunque scrittura di accordo di usufrutto, anche vincolante, purchè ritenuto vantaggioso e non preclusivo delle attività di libera pubblicazione di articoli scientifici, di settore e/o commerciali. Qualunque attività divulgativa non sarà comunque mai tesa alla concorrenza sleale e/o alla rivelazione di materiale oggetto di riservatezza industriale.

Le tecnologie

La piattaforma viene creata lungo due fasi propedeutiche, quindi successive, la prima della quale ha lo scopo di creare un prototipo funzionale di base (**PFB**) che sia sottoposto ad una fase di beta-testing per la validazione. CAIMANS nasce in fase di progetto come una piattaforma perché di per sé tutta la tecnologia software sarà utilizzata per creare una pletora di stumenti di *Middleware* proprietari.

Si consideri l'arco di flusso seguente :

a) **D-DBS → AKB ← IRS**

La esigenza di massima modularità della piattaforma è determinata dalla massima flessibilità nell'utilizzo futuro sia per ciò che sono i flussi testuali in ingresso (virtualmente qualunque fonte) e sia per ciò che riguarda le inferenze e le *queries* dell'output. Da qui la connotazione di *Middleware* per l'intera PPCA.

La PPCA essenzialmente inferenzia su una pletora di *Training Set* (**TSs**, vedi avanti) per creare e popolare in modo ricorsivo e quindi alimentare in modo progressivo una AKB

CAIMANS

Schema fattuale del progetto preliminare

White sheet

Il prototipo funzionale di base

Per intraprendere la costruzione del livello di prototipo funzionale della AKB si richiedono 3 elementi

1. Repertori di training sets
2. Software di servizio e sviluppo proprietario per la PPCA (Es. VB6, D-COM, ActiveX e ass-DLL)
3. Software operativo e mantenimento facilities (EMS, MS-SQL Ent. Manager)

Con il termine prototipo funzionale ci si riferisce all'assetto semi-lavorato di una KB che sia stata istruita con un numero preliminare di *training sets* sufficienti a verificarne : *performance*, congruità ed efficienza.

L'assetto delle risorse HW per il disegno, l'architettura e la messa in opera della piattaforma fino al rilascio della versione alfa, sarà fornita dalla consulente. Prima di procedere con la fase dei Beta test, e quindi dopo la revisione critica della figura del PM, si dovranno operare degli investimenti di minore entità per garantire tutte le licenze di proprietà intellettuale e i tools richiesti dalle scelte di *porting*.

La alta fattibilità del progetto è dovuta alla iniziale e immediatamente disponibile fonte di valore aggiunto di una Banca Data Euristica (**BDE**) di contenuti editoriali storicizzati. Disponendo di un bagaglio di esperienza formalizzata in una entità relazione editoriale con due livelli di chiave 1:1 e 1: molti per alcune decine di *t-ple* di classificazioni , essenzialmente la consulente potrà dare per acquisita la base euristica di una classificazione già filtrata dall'esperienza di giornalisti che hanno operato.

Oltre che su base euristica, la BDE è stata utilizzata su criteri addizionali di classificazione basati sulla esperienza tematica e/o di settore. Con questo bagaglio implicito, la BDE è la fonte migliore e più appropriata di estrapolazione dei *Taining Sets* per la istruzione deella AKB. Come sopra anticipato questo implica l'uso del termine suffisso "Associata" per la base di conoscenza KB che evidentemente riflette in modo "meta-cognitivo" la rappresentazione della macchina vettoriale di classificazione automatica derivante dalla banca dati storica.

Ovviamente, questa propedeutica fase di classificazione sfrutta come "*imprinting*" di un "*gold-standard*" i contenuti conosciuti, ma successivamente si renderà automaticamente reiterabile (*Daemon*) un motore che possa asservire un IRS semantico (S-IRS) in grado di generare nuovi ordini di classificazione e categorizzazione

Il setup iniziale DB Server

L'ordine indicativo delle operatività è vincolato agli stessi passi del disegno della PPCA. Pertanto la successione dei periodi è certa mentre la loro durata è indicativa per ovvie ragioni di gestione e conduzione di progetto secondo quanto realmente svolto dal PM

Lo schema sotto riassunto può ovviamente essere elongato su diverse scale di tempi in caso di forza lavoro addizionale. Ad esempio, i tempi potrebberò essere semplicemente dimezzati per la fase di linearizzazione e normalizzazione delle banche dati dei *training set* se soltanto si ipotizzasse una settimana/uomo di una figura di assistente DBA.

Propedeutico alla costruzione del gruppo di TS di disponibilità di un data base di training e test (**DBTT**). Tale formato verrà fornito alla consulenza dal PM per essere strutturato sotto una formato di studio transiente del servizio MS-SQL Server 2008.
Il servizio risiede nella macchina di sviluppo ma potrà essere pubblicato su una porta custom via policy proprietaria di DMZ/DHCP table gestita nell'accesso da WAN della intralan della consulente.
Questo essenziale costrutto sistemistico ha anche una ragione funzionale in quanto tutti i moduli sw di sviluppo della piattaforma potranno operare in modo parallelo e concorrenziale anche via EXTRANET. Evidentemente questo porterebbe ad un disegno vantaggioso di scala nel caso di molti *clients* di appoggio che dividano la computazione semaforica della PPCA.

La fase di configurazione del DBTT dovrebbe impegnare circa tre giorni, intendendo con ciò incluso anche il tempo per il quale congiuntamente l'archivio originale può essere raggiunto e utilizzato dal Flow-Controller (FC) remoto come servizio di tunnelling per il PM.

Bozza di flussogramma della PPCA

Al tempo di questo scritto non si è ancora formalizzato alcun tipo di contratto di fornitura nè di progetto, tuttavia ci sono stati alcuni incontri di *brainstorming* tra la consulente e il PM. Durante le fasi di ideazione e di scambio di specifiche/richieste, sono emersi alcuni elementi di valutazione che con buona approssimazione definiscono le funzionalità dei moduli portanti della piattaforma CAIMANS
In allegato a questo prospetto viene fornito il draft originale redatto dalla consulente in sede di primo incontro di

fattibilità con il PM (**Allegato AA1**)

Una fase propedeutica di lavoro presume l'uso di una classe di TSs già disponibili al tempo della partenza dei trattamenti. Ovviamente i trattamenti automatici di contenuto sfruttano in entrata la stringa testuale del campo BLOB del singolo arcitolo della DBTT.

L'inizio di processo dell'algoritmo di classificazione segue lo schema

b) **CTR → ACAP → SCSL → LFER**

dove

CTR : Contenuto testuale di Riga
ACAP : Algoritmo di Classificazione Automatica proprietaria
SCSL : Strato Categorico Semi Lavorato
LFER : Livello Fisico Entità Relazione

Più in esteso, il testo originale prelevato in ASCII dal tracciato fisico del singolo record viene passato alla cascata delle classi funzionali dell' algoritmo di classificazione che tipicamente tratta in memoria volatile lo stampo dell'oggetto categorico in tutte le sue componenti gerarchiche. Questo stadio della conoscenza viene anche individuato come Categorizzazione del Semi Lavorato. Infine l'astrazione del semi-lavorato, viene persistito nel modello di classificazione a livello fisico di struttura nel data base che formalizza la base di conoscenza.

Lo strato ACAP

Pivotale a tutto il progetto si trova l'algoritmo di classificazione automatica proprietaria. Oltre che per le ovvie ragioni legate alla proprietà intelllettuale di usufrutto della meta conoscenza originata, l'algoritmo è di fatto la espressione formalizzata della identificazione unicoca e peculiare della proprietà della logica euristica di tutti i contenuti di tutti i repertori processati.
In ragione della bontà del fattore semantico discriminante (FSD si vedrà più avanti nel disegno progettuale), la bonta dell'algoritmo determinerà il successo della flessibilità di accesso alla banca dati da parti di un IRS.
Vediamo di seguito alcuni elementi di processo dell' ACAP a livello funzionale.

La prima strutturazione di un approccio manuale/euristico riguarda l'uso di un copioso numero di Dizionari e Vocabolari per la tesaurizzazione progressiva dei Filtri Automatici di Conoscenza (**FAC**). A titolo indicativo la

tabella T1 riporta le principali routine di FAC sui viene sottoposto il CTR

In realtà il CTR rappresenta un esemplare unico che poi è soggetto a *versioning* generazionale in ragione del tipo di processazione del FAC. Per molti tipo di calcolo e manipolazione di stringa, il CTR originale viene propagato generazionalmente in copie multiple e interposte. Alcune di questi testi clonati sono transienti solo ai fini di calcoli interposti, altri permangono a scopo di back-propagazione e *time-stamp* nella struttura definitiva del LFER.

Tutti i passi dell'algoritmo sono monitorati da logs di tracciato e di validazione.

Deployement piattaforma di servizio

La realizzazione della piattaforma di prodotto si poggia sulla fase di sviluppo originale che usa una piattaforma di sviluppo. Gli strumenti di sviluppo utilizzati sono forniti a costo zero dalla consulente che sfrutta delle scelte di

a) ottimizzazione dei tempi di applicazione,
b) flessibilità di disegno e
c) performance computazionale.

Considerazione di interfaccia : per ciò ce riguarda i moduli di processo e di interfaccia di service, sarà utilizzato l'ambiente di sviluppo MS Visual Basic 6 SP5. Questa soluzione risulta compatibile per la piattaforma ospite del MS Windows Server 2008 a 32/64 bit. A corredo per alcune esigenze specifiche di eleborazione si potrà ricorrrere anche allla piattaforma RAD di Delphi 6.0 basando pertanto l'ambiente di sviluppo sul linguaggio Borland T-Pascal. In questo secondo caso il wrapping funzionale delle routine potrà avvalersi non solo di oggetti COM ma anche oggetti VCL.

La piattaforma di disegno e sviluppo dell'ACAP per la processazione dela fonte primaria testuale, avrà essenzialmente

a) moduli di console
b) moduli di calcolo di CA
c) moduli di *utility*
d) moduli di *maintenance* e supporto

In tutte le tipologie di interfaccia si ipotizzano applicativi verticali (processi segregati a 32 bit) che potranno condividere uno o più File Server, con DBS locali, intranet e/o remoti, e che disporranno di una nomenclata pletora di file di pilotaggio (INI stamp).
Tutti gli applicativi avranno funziolità **GUI** e **CLI**, cosi da poter essere monitorati e/o distribuiti anche su *thread* di sistema paralleli evitando però la duplicazione di oggetti ActiveX/DLL ActiveX e più in generale COM+ nella memoria del sistema.

Considerazioni di performance : gli elementi maggiormente critici nella fase di *deployement* e sviluppo della PPCA riguardano i due momenti di accesso al record con il campo testuale per la sua estrazione e le funzioni di processazione del testo di contenuto grazie alle routine di manipolazione di stringa.
Nel primo caso l'accesso è certamente ottimizzato perché il *driver* VB6 dal lato connettore DB viene svolto da un oggetto OLE-DB / ADO.NET proprietario dedicato al MS.SQL Server. Evidentemente non è rilevante quale è il linguaggio di programmazione dell'algoritmo visto che l'accesso ai dati per la estrazione del BLOB avviene per tutti a livello di sistema ed è *custom* per il DB server.
In termini di estrema semplificazione, che si stia costruentdo una IDE in Java, in C++, in Pyton, in Delphi o in VB6, il driver OleDB è per tutti lo stesso e nel momento in cui viene passata per riferimento la stringa contenente il campo testuale (ASCII) dell'articolo considerato, la *Memory Mapped Image* (MMI) della stringa è parte dell'oggetto in memoria dell' *In Memory Record* (IMR).

Performace in test e sviluppo : gli aspetti di *performance* in ambiente di *testing* riguardano soprattutto la processazione della stringa. In questo ambito cruciale risulta il *task* della manipolazione di stringa.

Tipicamente in informatica la tecnologia più consolidata e al contempo più definita nel trattamento aspecifico delle stringhe è quella delle Regular Expression (**RegEXP** o più brevemente **RE**).
Virtualmente nessuna piattaforma e/o marchio di produzione di un certo rilievo, ha mai potuto prescindere da una propria implementazione di un motore per le RE. In particolare si cita il linguaggio PERL come interfaccia più ottimizzata per le RE.
La consulente, per il progetto CAIMANS ha scelto un set di librerie DLL costruite e compilate in linguaggio Assembler. Il prodotto commerciale si chiama STAMINA32 ed è stato usato dalla consulente sin dal 1998.

La peculiarità di queste librerie sta prima di tutto nel fatto che sono state costruite per interfacciarsi eletticamente con l'ammbiente VB6, in secondo luogo per la velocità di computazione assolutamente impareggiata nell'ambito del trattamento delle stringhe, e degli Array LILO/FIFO dal lato di programmazione wrapping della DLL.

Da ultimo, non trascurabile, le librerie sono state acquisite come Royalty Free e sono distribuibili senza vincoli di usufrutto dalla consulente. Ovviamente a fronte di accordi in essere da definire.

Performace in produzione : essenzialmente la struttura a modulare di processo degli applicativi di Back-end risolve la problematica delle performace di processo con la scalabilità hw; come abbiamo visto al paragrafo precedente a questo si aggiunge l'uso di DLL scritte in assembler di proprietà della consulente ed elettivamente interfacciate con l'ambiente di sviluppo VB6 SP5.
Esiste tuttavia un ultimo livello di ottimizzazione delle performance che può essere utilizzato per la fase di produzione. Per il tempo in cui tutte le fasi di Beta test siano state saggiate e la bontà dei processi di PPCA è stata verificata, esiste la possibilità di una fase di Porting per cui sarà sussificiente traghettare la struttura funzionale delle applicazioni verticali verso sorgenti C++ e/o C# a seconda della onerosità e della valutazione costo beneficio.

In tutti i casi le scelte di porting prescinderanno dalla fruibilità futura dell'intera AKB dal lato web. Eventuali interfaccie verso la base di conoscenza saranno progettate secondo un disegno del tutto dissociato dalla piattaforma di back-end.
Una ultima considerazione sulla scelta di porting nella piattaforma MS Visual Studio riguarda la disponibilità di un Assembly e un oggetto distribuito di libreria per la Regular Expression, già ottimizzato per la chiamata da C# e Vb .NET.

Famiglie di categorizzazione

Le attività di Text Categorisation sono piuttosto standardizzata, ma nella nuova PPCA che si intende costruire, saranno introdotte anche le Famiglie di Categorizzazione (**FDC**). Alcuni esempi sono stati accennati nela bozza del disegno di cui in Allegato 1 e sono

a) Materie
b) Ambiti
c) Domini

Queste super classificazioni non sono da interdersi come legate gerarchicamente ma come proprietà di uno stesso classificatore di indice associato ad ogni singolo articolo. La lista sopra indicata è solo indicativa e non esaustiva. La esigenza di scrutinare nuovi ordini di classificazione deriva dalla intenzione di poter nel futuro traversare la base di conoscenza secondo interessi multipli di contesto.

L'analisi contestuale formalizzata da diversi punteggi di TF caratteristici di classificatori diversi, costituirà la proprietà semantica che si intende aggiungere al calcolo deterministico della FTS.

La nomenclatura di classificazione settoriale (By Sector Classification BSC) è utilizzata da molti Network di Public Domains & Institution. Si pensi a EUROSTAT che gratuitamente mette a disposizione il proprio Guide Lines di classificazione di settore ad esempio per tutto ciò che concerne le statistiche dei Patents. Le Figure 1°, 1b e 1c riportano alcune immagini di riferimento tratte dal portale ufficiale della Comunità Europea.

CAIMANS

Schema fattuale del progetto preliminare

White sheet

Tabella T1 - principali routines di FAC sui viene sottoposto il CTR per il TF

Clean-up	*Pulizia articoli, elementi sintattici,*
Nationalisation localisation	*Slangs, forme contratte tipiche*
Inverse / reverse bias	*Duetti, triplette reversali classificate*
Stop-words Analisys	*Es. Giorni della settimana, acronimi non classificati, fuga di controllo*
Duplicate /redundants	*Duplicazioni di chiavi note, fuga di controllo*
Syntax calibration	*Virgolettati, frasi e frasari classificati come topics*
Normalisation FTC	*Punteggi discreti e deterministici distanza fenetica CA*
Indicizzazioni categoriche	*Schemi di indicizzazione degli enti e alberi BTREE di classificazione*

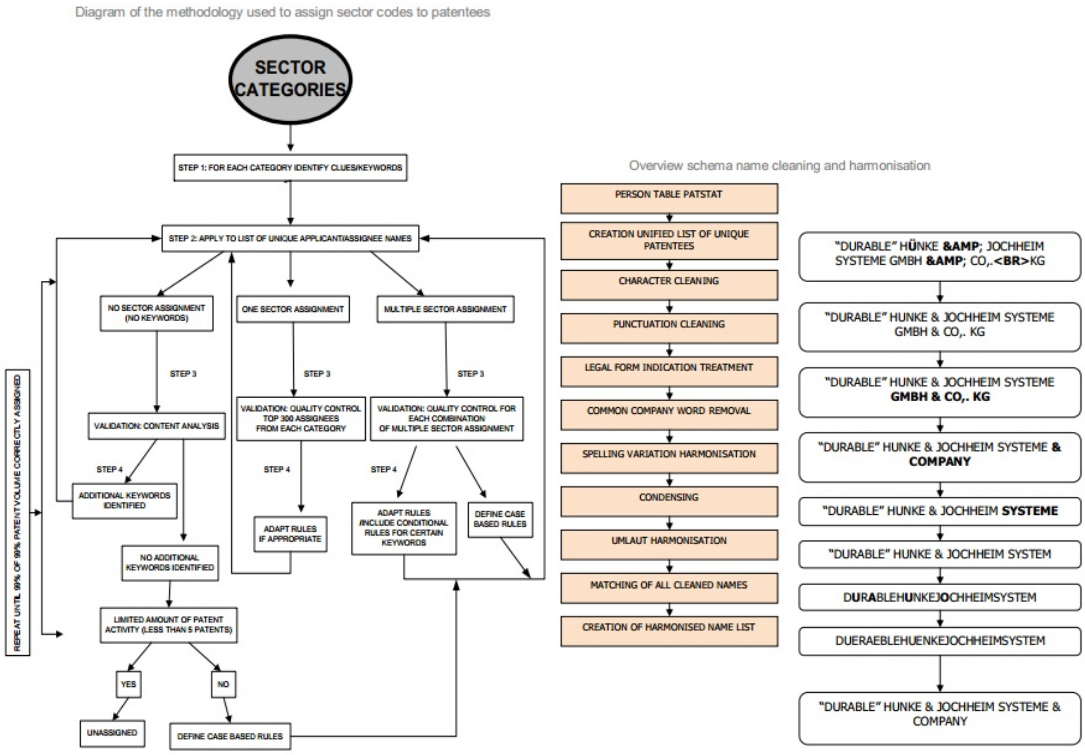
TFC : Terms Frequency Calculation ; CA : Classificazione Automatica

CAIMANS

Schema fattuale del progetto preliminare

White sheet

Figura 1A/1B e 1C : EUROSTAT 2011 Classification Standard - diagramma dell' algoritmo, harmonisation tables



Examples of keywords/clues used to identify patentee sectors

Sector	Keywords
(1) Individual	"*DIPL.-ING.*"; "*PROF.*"; "*DR.*"; "*DECÉDÉ*"; "*DECEASED*"; "*DIPL. ING.*"; "*PH.D*"; "*DIPL.-GEOGR.*"; "*ING.*"; "*ÉPOUSE *"
(2) Private Enterprise	"* SA *"; "*S.R.L*"; "*HANDELSBOLAGET*"; "*INC."; "*LTD."; "*S.A.R.L"; "* BVBA *"; "*S.P.R.L.*"; "*NAAMLOZE VENNOOTSCHAP*"; "*AKTIEBOLAG*"
(3) Public and Private Non-Profit	"*GOUVERNMENT*"; "* MINISTRO*"; "*INSTIT*"; "*INSTYTUT*"; "*FONDATION*"; "*FOUNDATION*"; "*CHURCH*"; "*TRUST*"; "*KENKYUSHO*"; "*STIFTUNG*"
(4) University	"*UNIVERSI*"; "*UNIV.*"; "*COLLEGE*"; "*SCHOOL*"; "*REGENTS*"; "*ECOLE*"; "*FACULTE*"; "*SCHULE*"; "*UNIVERISTY*"; "*UNIVERSTIY*"
(5) Hospital	"*HOSPITAL*"; "*MEDICAL CENTER*"; "*MEDICAL CENTRE*"; "*ZIEKENHUIS*"; "*CLINIQUE*"; "*NOSOCOMIO*"; "*CLINICA*"; "*POLICLINICA*"; "*HOPITAL*"; "*HOPITAUX*"

CAIMANS

Schema fattuale del progetto preliminare

White sheet

Allegato AA1 – *Bozza preliminare del disegno di flusso per la creazione di un network modulare di processi di CA*