LIVEMEMORIES

Memorie Digitali Attive di Vita Collettiva

WWW.LIVEMEMORIES.ORG

PROGETTO AMMESSO AL FINANZIAMENTO DEL BANDO GRANDI PROGETTI 2006
CON DELIBERAZIONE DELLA GIUNTA PROVINCIALE N. 686 DEL 18 MARZO 2008

Coordinatore: Fondazione Bruno Kessler (FBK)
Via Santa Croce, 77 - 38100, Trento (Italia)
Coordinatore Scientifico: Bernardo Magnini

Soggetti Partecipanti: Università di Trento (UNITN)
Via Belenzani, 12 - 38100, Trento (Italia)
Coordinatore: Massimo Poesio

Università di Southampton (SOTON)
University Road, Southampton SO17 1BJ (UK)
Coordinatore: Wendy Hall

Contatto: Bernardo Magnini (magnini@fbk.eu)

Inizio Progetto: 1 ottobre 2008
Durata Progetto: 36 mesi

Sommario

1. Introduzione

Nell'era digitale le testimonianze del passato e del presente si moltiplicano: da un lato sono in corso grossi sforzi per digitalizzare materiali su supporto analogico, dall'altro sono ormai diffusi strumenti per la raccolta di memorie in diversi formati (es. immagini, video, testi), con dispositivi a basso costo (es. macchine fotografiche digitali, cellulari). Questa ricchezza di dati, congiunti con le nuove opportunità di condivisione tramite piattaforme Web (es. Flickr, blogs), aprono potenzialità completamente nuove di fruizione del ricordo e di partecipazione all'esperienza collettiva della memoria.
Questa enorme quantità di dati sono però scollegati, incompleti e inconsistenti, distanti nel tempo e nello spazio, ed espressi su media diversi. Occorrono nuovi strumenti per "farli parlare" tra di loro. La sfida scientifica di LiveMemories sta nello sviluppo di metodi per interpretare in modo automatico il contenuto di tali frammenti per trasformare l'enorme quantità di dati multimediali oggi disponibili, tasselli che, come in un immenso puzzle digitale della memoria collettiva, devono essere integrati, completati, adattati per ricostruire le figure del nostro passato, in "memorie vive".
Immaginate il portale creato da una scuola al fine di costituire una storia della comunità attraverso la condivisione di testi, foto e video, in cui uno studente inserisce una fotografia con la nota "Mario Rossi con i compagni della 3A al liceo Da Vinci nel 1972". In una memoria "viva", la fotografia e il commento saranno trasformati da informazione statica a contenuto attivo grazie all'interpretazione di testo e immagine, ed al collegamento con altri dati esistenti sulla piattaforma—per esempio, dati sugli studenti della 3A nel 1972, loro foto attuali, etc.
La sfida per LiveMemories verrà dalla natura stessa dei dati che costituiscono le memorie digitali, le quali (i) devono essere compilate da fonti eterogenee; (ii) sono dinamiche, cioè soggette a cambiamenti nel tempo; (iii) sono spesso contraddittorie o incomplete; (iv) devono essere adattate ai bisogni dell'utente.
Per far fronte alle sfide scientifiche, il progetto sfrutterà le competenze dei partecipanti nelle aree delle Tecnologie del Linguaggio, Knowledge Management e Web Science. Più specificamente, l'obiettivo è quello di migliorare lo stato dell'arte nelle aree di (i) estrazione di contenuti da fonti eterogenee, (ii) integrazione di contenuti attraverso l'applicazione di tecniche di ragionamento su larga scala, (iii) presentazione di contenuti attraverso la realizzazione di viste adattive. Un risultato concreto e permanente sarà la Piattaforma di Gestione dei Contenuti, attraverso la quale sarà possibile monitorare e memorizzare giornalmente i contenuti diffusi dai media trentini.
L'efficacia ed impatto di LiveMemories sarà valutata tramite un'iniziativa di raccolta di memorie collettive nella quale sarà coinvolta una comunità locale. Sarà reso disponibile al pubblico un portale per la raccolta, la gestione, l'integrazione e la fruizione di memorie collettive multimediali (testi, immagini con annotazioni scritte, filmati, mappe) provenienti da varie fonti (associazioni sul territorio, giornali ed emittenti locali) con il coinvolgimento anche di famiglie e di singole persone con l'obiettivo di mettere a disposizione frammenti di storia collettiva. La visibilità del progetto sarà anche assicurata da una mostra LiveMemories organizzata per far vedere come i progressi tecnologici che sottostanno all'idea delle memorie come contenuti digitali attivi possano offrire nuove modalità di vivere le memorie collettive del nostro passato, con un forte impatto sociale.

2. Obiettivi Generali

Da un punto di vista scientifico e tecnologico, LiveMemories punta a: sviluppare tecniche di estrazione del contenuto in grado di trattare enormi quantità di dati multimediali, ponendo le basi per una Piattaforma di Gestione dei Contenuti per il Trentino; usare i dati così estratti per realizzare nuove forme di riassunto e classificazione dei dati in una nuova generazione di memorie digitali che sono 'vive'; e trasformare la creazione di tali memorie in un'attività comunitaria sul Web. Raggiungere questi obiettivi ci darebbe grande visibilità nell'ambito della Web Science Initiative, delle memorie digitali, e del Web 2.0, grazie anche al coinvolgimento di Southampton. Ma LiveMemories non dovrà avere solo un impatto scientifico, ma anche sociale, tramite la raccolta, analisi e conservazione di memorie digitali del Trentino; il facilitare ed incoraggiare la conservazione di memorie collettive; ed il conseguente sviluppo di nuove forme di comunità ed arricchimento sociale e culturale.

3. Stato dell'Arte

Presentiamo lo stato dell'arte nelle Digital Memories, l'area focale del progetto, e in tre aree di attività scientifica: estrazione di contenuto da multimedia, integrazione di contenuti e loro presentazione.

MEMORIE DIGITALI E COLLETTIVE
La rivoluzione digitale crea nuove opportunità per raccogliere e mostrare le memorie del passato. Le memorie digitali rendono artefatti storici (per es., libri troppo antichi per essere maneggiati) accessibili a tutti e offrono ai professionisti tempi di ricerca ridotti. Per la commissione europea le biblioteche digitali sono un aspetto chiave dell'iniziativa i2010. Tra le iniziative attuali nell'area menzioniamo:

- MEMORIA PER IL TRENTINO (www.trentinocultura.net/memoria/hp_memoria.asp), progetto dedicato a scoperta, salvaguardia e promozione dell'eredità storica, artistica e culturale, specie in relazione all'ultimo secolo, in cui la comunità locale ha conservato la propria identità malgrado i cambiamenti avvenuti.
- MALACH (malach.umiacs.umd.edu) in cui tecnologie di riconoscimento automatico del parlato e traduzione assistita migliorano l'accesso agli archivi di storie orali videoregistrate della Shoah Visual History Foundation.
- MEMORIES FOR LIFE (www.memoriesforlife.org) indaga su comunanze tra memoria umana e automatizzata per sviluppare tecnologie per immagazzinare e recuperare più efficientemente le memorie in domini personali, sociali e di lavoro.
- E-MAIL BRITAIN, partita nel mag.'07, è un'iniziativa della British Library che intende raccogliere dal pubblico 1M di mail su diversi argomenti.

La preservazione digitale della memoria diventa più rilevante con il web 2.0—il crescente interesse verso piattaforme basate su web per condividere e creare insieme informazione,

come Wikipedia, Facebook, o Flickr. La partecipazione pubblica nella creazione della memoria può portare a scoprire documenti di valore non disponibili alle istituzioni.
L'interesse nelle memorie digitali e condivise è evidenziato dal numero di eventi sul tema, come i workshop CARPE che trattano di raccolta, recupero, organizzazione, ricerca, e questioni legali e di privacy, nell'ambito dell'archivio continuo e della gestione di tutte le esperienze personali relative ai media, o CIRN (Community Informatics Research Networks, ottobre 2006) dal tema "Constructing and sharing memory: Community informatics, identity and empowerment", focalizzatosi sulla costruzione della memoria, su quale sia il ruolo delle comunità informatiche nello sviluppo di nuovi mezzi per catturare la memoria privata e quella pubblica e su come le tecnologie dell'informazione e della comunicazione aiutano le comunità.

ESTRAZIONE DEI CONTENUTI
LiveMemories si propone di estrarre contenuto digitale da tre sorgenti: collezioni di testi, trascrizioni automatiche del parlato e database di immagini.

Estrazione del Contenuto dai Testi
Diversi strumenti ad elevate prestazioni per il POS tagging e parsing sono ormai di comune dominio almeno per la lingua inglese; a sua volta questa disponibilità ha permesso di applicare l'elaborazione semantica dei testi su larga scala (estrazione delle entità rilevanti di un testo, l'individuazione automatica della loro coreferenza, estrazione di relazioni e popolazione automatica di ontologie). Alcuni progetti degli USA, come MUC, ACE, AQUAINT, e recentemente GALE, hanno reso disponibili grandi risorse annotate e introdotto tecniche di valutazione quantitativa. Per quanto riguarda le coreferenze all'interno di documenti (IDC), questo ha portato allo sviluppo dei primi modelli di machine learning su larga scala che utilizzano queste risorse annotate [9.12, 9.13] e allo sviluppo dei primi strumenti per l'individuazione di coreferenze all'interno dei documenti. Piu` recentemente, il sistema ELKFED/BART sviluppato al workshop ELERFED (www.clsp.jhu.edu/ws2007/groups/elerfed/). Nel campo dell'estrazione di informazioni, il lavoro svolto come parte dell'iniziativa ACE e in ELERFED ha mostrato che buoni risultati possono essere ottenuti estraendo informazioni dalle notizie oon metodi supervisionati, in particolare Support Vector Machines (SVMs) e Kernel Methods [10.1], ma questi metodi semi-supervisionati sono piu` efficaci con testi meno formali.
L'enfasi attuale riguarda il miglioramento delle prestazioni e dell'utilita` degli strumenti esistenti:

- Creando grandi corpora e sviluppando tecniche che fanno a meno della necessità di annotazioni su larga scala (active learning, metodi weakly supervised).
- Considerando progettazione di tecniche di pre-elaborazione migliori (un aspetto cruciale e ancora spesso sottostimato).
- Incorporando tecniche di estrazione automatica del lessico e del senso comune, oltre a quelle tradizionali 'di superficie'.
- Sviluppando metodi di Machine Learning (ML) migliori, capaci di sfruttare queste sorgenti di informazione più avanzate (e.g. migliori funzioni kernel).
- Sviluppando rappresentazioni di relazioni più complesse, ad esempio, con modifiche temporali (ad esempio John Doe era CFO dell' ACME dal 2001 al 2005);
- Studiando in maggiore profondità metodi automatici per Textual Entailment Recognition [9.12], un tipo robusto di inferenza testuale basato su schemi acquisibili automaticamente dai corpora.

Estrazione del contenuto dal parlato
Il riconoscimento automatico del parlato ha raggiunto un livello tale che permette di gestire un grande vocabolario ed il riconoscimento del parlato continuo indipendente dal parlante. Malgrado la consapevolezza del livello della tecnologia sia bassa, la dettatura dei testi in ambienti professionali è spesso supportata dal riconoscimento automatico del parlato. La ricerca in questa area si è recentemente focalizzata sulla sfida successiva, ossia il riconoscimento del parlato da parlanti non cooperativi. In altre parole, gli oratori non parlano con l'intenzione di essere compresi da un sistema automatico, e.g. telegiornali, conversazioni, discorsi e cosi via. Tipici scenari applicativi sono la trascrizione dei meeting o l'indicizzazione di biblioteche digitali audiovisive.
Inoltre, gli analisti del riconoscimento automatico del parlato, hanno considerato anche la sfida di estrarre informazione rilevante. I problemi aperti riguardano il riconoscimento automatico, prima dell'elaborazione testuale, dell'inizio e fine delle frasi, delle parole in maiuscolo o minuscolo e delle intonazioni.

Estrazione del contenuto da immagini
Diversi passi in avanti sono stati fatti anche nel campo dell'estrazione automatica del contenuto da immagini digitali e video. In molti casi, sistemi specifici producono un'alta prestazione, particolarmente quando assunzioni semplificative possono essere fatte riguardo il dominio di applicazione. Si considerino tre esempi chiave:

- La semplice progettazione di sistemi con tecniche simili a quelle usate nei lettori di codici a barre.
- Tecniche per il riconoscimento dei volti (e altri sistemi biometrici visuali) sono sufficientemente affidabili per usi commerciali diffusi.
- Diagnosi mediche da immagini (e.g. mammografie) è un'altra area di studio importante.

Il dominio di LiveMemories presuppone forti richieste in termini dell'estensione dell'applicazione prevista.
Non solo molteplici contenuti saranno presenti, ma ci saranno problemi ulteriori con una larga varietà di formati, linguaggi, vocabolari, e cosi via. Lo sfruttamento di ontologie e altre tecnologie associate con il Semantic Web è importante per questo problema di ricerca. Tecniche di navigazione e indicizzazione assieme a metodi per la conversione di metadati in RDF ben strutturato possono essere usati con questi contesti per supportare interrogazioni di database multimediali.
Inoltre un grande ammontare di dati per le immagini digitali è disponibile a basso costo per mezzo della raccolta automatica dei metadati che usano formati come l'Exchangeable Image File Format (Exif).
Questo rappresenta informazioni chiave come la posizione GPS dell'immagine, il tempo e l'obiettivo (focus). Le diverse combinazioni dei riconoscitori di caratteristiche dirette dell'immagine (e.g. riconoscitori di linee) con l'informazione associata (e.g. notizie online o previsioni meteo), mostrano che l'estensione di questa area di ricerca su supposizioni intelligenti a proposito delle immagini è piuttosto ampia.

INTEGRAZIONE DEI CONTENUTI
Integrare i contenuti estratti significa riconoscere automaticamente la co-referenza tra entità, ed assegnare ad ognuna di esse un contesto geografico e temporale. Tutto ciò dovrà

essere memorizzato in ontologie e dovranno essere forniti un insieme di strumenti di ragionamento adeguati su tali ontologie.

Coreferenza intra-documenti
Nonostante, l'interesse sul tema della co-referenza tra menzioni in documenti diversi sia piuttosto recente [9.14], negli ultimi anni si sono sviluppati una serie di contributi importanti. Tale sforzo è riconducibile alla grande importanza di questo topic in applicazioni sia nella pubblica amministrazione che nelle industrie private. Più in particolare, c'è stato un grande interesse in una forma più semplice di disambiguazione di entità conosciuta con il nome di "Web entity", come ad esempio il task "web people" di SEMVAL [9.15], e il "spock challenge" (challenge.spock.com) emerso nel "Web People task" di SEMVAL i sistemi allo stato dell'arte sono basati su clustering di descrizioni di entità. Queste ultime contengono un mix di informazioni su locazione, entità e relazioni tra entità. SEMVAL ha inoltre mostrato che il clustering ed in particolare il criterio di terminazione sono cruciali. Infine, la competizione "Spoke" ha evidenziato la necessita di metodi in grado di lavorare con grandi masse di dati. Di conseguenza, recenti sviluppo si sono concentrati sul miglioramento delle tecniche di clustering e sulla sperimentazione di tecniche robuste per l'estrazione di informazioni da testi. (ad es, i risultati con ELERFED).
La maggioranza delle ricerche qui menzionate riguarda l'Inglese; alcuni progressi sono stati fatti anche per Tedesco (ad es. Versley) e Spagnolo (Ferrandez), ma pressoché nulla è stato fatto per l'Italiano (con l'esclusione di [9.16]).

Servizi di ragionamento complesso
Nell'ambito della rappresentazione della conoscenza e ragionamento automatico (KRR) sono stati sviluppati un insieme di formalismi e strumenti di ragionamento che sono usati anche in applicazioni reali. La logica descrittiva (DL) [9,7] costituisce attualmente il paradigma più diffuso e sistemi come Pelle, Racerpro e Fact++ sono disponibili per la maggioranza delle applicazioni nel semantic web. Modularizzando la conoscenza (cioè organizzando la conoscenza in 'piccoli' moduli) si è anche raggiunto un certo livello di scalabilità. Il sistema DRAGO ad esempio fornisce un supporto al ragionamento per conoscenza modulare [9.8]. Sono stati fatti alcuni studi per l'estensione della DL per rappresentare conoscenza incerta [9.9]. I database temporali, invece, forniscono un sistema di background molto maturo per la rappresentazione della conoscenza legata al tempo. Ma più recentemente sono stati fatti del lavori (ad es. OWL-MeT ermolayev.com/owl-met/) che estende il linguaggio di DL per trattare costrutti temporali e proprietà metriche del tempo. In modo simile, sono state proposte estensioni di DL per trattare conoscenza incerta. La rappresentazione delle attitudini proposizionali (ad esempio credenze e intenzioni, idee, e opinioni) è stata molto studiata in KRR ma non sono stati sviluppati degli strumenti affidabili e maturi. Infine non c'è pressoché niente per la rappresentazione di conoscenza che coinvolge tutti gli aspetti in un unico sistema.

Popolamento di ontologie
L'apprendimento e il popolamento di ontologie con dati derivanti da documenti multimediali è un tema di ricerca attualmente molto importante. L'apprendimento di ontologie, ha come obiettivo quella della costruzione (semi) automatica delle stesse partendo da dati semi-strutturti o da dati testuali. Il popolamento di ontologie invece ha come obiettivo quello di "riempire" un'ontologia esistente con individui (= oggetti) e relazioni tra individui che sono estratte automaticamente da documenti multimediali.
Text-to-Onto, Web->DB sono due esempi di strumenti per l'apprendimento di ontologie, basati sul Formal Concept Analysis e reti Bayesiane, ma non sono gli unici [9.11] contiene

una descrizione e valutazione di strumenti di apprendimento di ontologie. Esistono anche svariati strumenti di popolamento di ontologie, principalmente basati su tecniche di analisi del linguaggio naturale. Ad esempio all'interno del progetto Ontotext l'FBK ha sviluppato un metodo per la popolazione automatica di ontologie che descrivevano persone.

PRESENTAZIONE DEI CONTENUTI

Diverse tecniche permettono di accedere a vaste collezioni multi-dimensionali di contenuti.

Queste includono:

- Classificazioni gerarchiche (HC): alberi con un unico nodo come radice, nei quali ai nodi sono assegnate etichette in linguaggio naturale (categorie). La categoria di un nodo figlio rappresenta una sotto-categoria o una specificazione della categoria del nodo genitore. Le HC rappresentano un modo naturale per organizzare e accedere a contenuti e vengono utilizzate nelle biblioteche, nella ricerca sul web, nel commercio elettronico, nella gestione di conoscenza personale, ecc.
- Clustering [9.1]: una tecnica che si basa su misure di similitudine per raggruppare automaticamente degli elementi. Un esempio di clustering di risultati Web si può vedere su Clusty.com.
- Classificazione "Faceted" [9.1]: consente la descrizione di una raccolta attraverso diverse caratteristiche ortogonali comuni a tutti gli item. Gli utenti possono generare viste diverse applicando le caratteristiche in un ordine arbitrario definito dall'utente.
- Navigazione: permette di identificare categorie del contenuto che si riferiscono ad una determinata categoria ma non sono direttamente collegate ad essa. La navigazione è applicabile sia all'interno di una singola classificazione gerarchica che in un insieme di classificazioni gerarchiche. Permette di saltare da un nodo di un ramo a un nodo di un altro, come per es. @-links in dmoz open directory project [9.4].
- Ricerca Sintattica (SyS): un approccio alla ricerca di contenuti effettuata confrontando parole chiave e/o valori di attributi, messi a disposizione dagli utenti, con il corpo e/o gli attributi degli elementi di una collezione. Nella ricerca sintattica il matching è risolto valutando le relazioni tra le stringhe.
- Ricerca Semantica (SeS) fa riferimento all'uso della semantica nella costruzione della query, nel processo di ricerca e nella visualizzazione dei risultati della ricerca [9.2]. Nella costruzione della query, la semantica è attivata mediante l'uso di vocabolari controllati, la disambiguazione della query e l'applicazione di costrizioni semantiche sul significato dei termini della query. Durante la ricerca, la semantica è implementata in termini di espansione della query, attraversamento di grafi, propagazione, e ragionamento RDFS/OWL. Nonostante la presenza in letteratura di circa 35 sistemi per la ricerca semantica, c'è una evidente mancanza di valutazione degli algoritmi di ricerca semantica, di valutazioni delle interfacce e delle API da parte degli utenti [9.3].
- Tecniche incentrate sugli aspetti sociali: negli ultimi anni il Web 2.0 ha ripreso approcci comunitari già proposti (almeno in parte) dal CSCW. Tra questi citiamo social filtering e recommendation system, folksonomie, tagging and social bookmarking, social navigation. Nuovi approcci coma la "human computation" [9.5] stanno nascendo in quest'area.

Riferimenti Bibliografici:

[9.1] Hearst, Clustering versus Faceted Categories for Information Exploration, Communications ACM 49(4), 2006
[9.2] Hildebrand et al., An analysis of search-based user interaction on the Semantic Web. Information Systems Center, INS-E0706, 2007
[9.3] Shvaiko et al., A Survey of Schema-based Matching Approaches. JoDS, IV, 146, 2005
[9.4] Glover et al., Using Web Structure for Classifying and Describing Web Pages, Proceedings of WWW-02, 2002
[9.5] van Han et al., Games with a purpose. IEEE Computer, 2006
[9.6] Giunchiglia et al., Semantic Matching: Algorithms and Implementation. JoDS IX, 2007
[9.7] Baader et al., The Description Logic Handbook. CUP, 2003
[9.8] Serafini et al., DRAGO: Distributed reasoning architecture for the semantic web. ESWC 05, 2005
[9.9] Straccia, Reasoning with Fuzzy Description Logics. JAIR14, 2001.
[9.10] Buitelaar, Ontology learning from text. Tutorial at ECML/PKDD, 2005
[9.11] Popescu et al., From Mention to Ontology: A Pilot Study. SWAP, 2006.
[9.12] Szpektor et al., Scaling Web-based Acquisition of Entailment Relations. EMNLP, 2004.
[9.13] Dagan et al., Direct Word Sense Matching for Lexical Substitution. ACL, 2006
[9.14] Bagga and Baldwin, Entity-based Cross-document Coreferencing Using the Vector Space Model. ACL, 1998
[9.15] Popescu and Magnini, IRST-BP: Web People Search Using Name Entities. SEMEVAL, 2007
[9.16] Delmonte et al., VENSES - a Linguistically-Based System for Semantic Evaluation. PASCAL, 2005
[10.1] Moschitti et al, Tree Kernels for Semantic Role Labeling. CL, 2008.
[10.2] Moschitti et al, Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification. ACL, 2007
[10.3] Moschitti and Zanzotto, Fast and Effective Kernels for Relational Learning from Texts. ICML, 2007
[10.4] Riccardi et al., NEEDLE: Next Generation Digital Libraries, AISV, 2006
[10.5] Tuffield et al., Towards the Narrative Annotation of Personal Information and Gaming Environments. Hypertext, 2005
[10.6] Domingue et al., Supporting Ontology Driven Document Enrichment Within Communities of Practice. Knowledge Capture, 2001
[10.7] Moschitti et al., Spoken Language Understanding with Kernels for Syntactic/Semantic Structures. ASRU, 2007

4. Risultati Preliminari conseguiti dai proponenti

Il progetto proposto costruisce sopra e integra i risultati di importanti esperienze di sviluppo tecnologico conseguite in precedenti progetti dai partner partecipanti.

ESTRAZIONE DEI CONTENUTI

Text Processing

Per quanto riguarda il cleaning dei dati e la raccolta di corpora, UniTN è stato un partecipante chiave nella campagna per la creazione del Wacky corpus e dell'iniziativa CLEANEVAL, creando anche Jboot, uno strumento per la raccolta automatica di corpora basati sul web. I-CAB, Italian Corpus Annotation Bank (FBK), è una raccolta di 500 articoli tratti dal giornale locale Adige, annotato semanticamente a diversi livelli, comprese menzioni di entità (sia descrizioni che entità nominate) e espressioni temporali.
UniTN vanta molti anni di esperienza nell'annotazione di coreferenze e relazioni sia per l'italiano che per l'inglese. FBK ha anche realizzato MultiWordNet, la versione italiana dell'inglese WordNet.
Gli strumenti realizzati includono TextPro, sviluppato da FBK, un set di strumenti per l'elaborazione dell'italiano che include POS tagging, lemmatizzazione e riconoscimento

delle Named Entity per l'italiano; strumenti per la rappresentazione e l'indicizzazione, basata sulla piattaforma Lucene, di corpora di grandi dimensioni. Fondamenta solide per un ulteriore lavoro su coreferenza intra documento sono state stabilite come parte del già menzionato workshop ELERFED all'Università Johns Hopkins, finanziato da US NSF e dal Dipartimento della Difesa statunitense, durante il quale lo strumento BART è stato sviluppato; è stato attualmente messo a disposizione per open source release tramite SourceForge. Tease è uno strumento per l'acquisizione di entailment patterns dal web (FBK). FBK è stato l'organizzatore di Evalita 2007 (evalita.itc.it), in cui si è effettuata una valutazione comparativa degli strumenti per il NLP dell'Italiano.

Speech processing.
FBK ha una lunga tradizione nello sviluppo di tecnologie di frontiera per il riconoscimento automatico del parlato. In particolare, la tecnologia sviluppata da FBK per la trascrizione automatica del parlato rintracciato in registrazioni audio è stata recentemente raffinata e validata nel progetto europeo TC-STAR.
Questa tecnologia include moduli per la partizione automatica delle registrazioni audio, la decodifica con grande vocabolario, l'adattamento al parlatore/ambiente, e il re-scoring basato su grafi di parole. Per lo sfruttamento commerciale della tecnologia sviluppata, una licenza d'uso esclusivo è stata rilasciata alla spin-off di FBK PerVoice.

Image processing
SOTON ha un'esperienza di molti anni nell'estrazione di contenuti in campo multimediale. Ad esempio: MAVIS, un'architettura all'interno della quale moduli per il processing associati con features estratte da media possono essere aggiunti e integrati; SCULPTEUR, che utilizza tecnologie Semantic Web per il reperimento di oggetti museali 3D.; MIAKT che utilizza SW e tecnologie basate sulla conoscenza per supportare e gestire analisi di immagini in campo medico; PHOTOCOPAIN, un sistema di annotazione di immagini semi-automatico, in grado di unire informazioni sul contesto con altre informazioni disponibili, in modo tale da generare annotazioni di contorno che l'utente può estendere ulteriormente oppure modificare.

INTEGRAZIONE DI CONTENUTI
Durante il workshop ELERFED sopra citato, molteplici sistemi disambiguazione di entità sono stati costruiti in collaborazione con Gideon Mann e colleghi dell'Università del Massachusetts a Amherst, e testati sui dati Spock. FBK è stato impegnato per alcuni anni nella ricerca su entity disambiguation sia come parte del progetto OntoText sia partecipando a SEMEVAL Web People Search nel 2007, durante il quale si è classificato al secondo posto.
Geo-time tagging. Il sistema Chronos, sviluppato da FBK, permette il riconoscimento e la normalizzazione di espressioni temporali per l'italiano e l'inglese. E` stato valutato dall'iniziativa Evalita, classificandosi al primo posto.
Ontology population. Nel progetto Ontotext finanziato dalla PAT, FBK ha sviluppato uno strumento chiamato ontology repository, in grado di supportare la gestione della conoscenza di base sotto forma di ontologie OWL, e fornire funzioni di ricerca e ragionamento da sfruttare per il popolamento di ontologie.
Nel progetto APOSDLE (finanziato dall'EU) che supporta l'e-learning, FBK ha sviluppato uno strumento per la specificazione di ontologie comuni attraverso wiki, e per la traduzione automatica in ontologie OWL.
Il sistema di ragionamento DRAGO, sviluppato da FBK, supporta il ragionamento distribuito tra ontologie integrate ed eterogenee.

PRESENTAZIONE DEI CONTENUTI
Sweb (http://www.dit.unitn.it/~knowdive) è un sistema per la gestione di dati e di conoscenza sviluppato presso l'Univeristà di Trento. Permette agli utenti di creare classificazioni gerarchiche, di popolarle con documenti di vario tipo, cercare immediatamente classificazioni e documenti, creare link per "navigare" tra le classificazioni e usare questi link per ricerca mediante browser o distribuita, condividere classificazioni, documenti e link con altri utenti in una modalità completamente controllata, ed eseguire altre operazioni.
Sweb è un sistema semantics-enabled in quanto possiede una conoscenza di base in grado di codificare fatti riguardanti un dominio specifico; questa conoscenza di base è utilizzata per supportare l'utente nei compiti di gestione dei dati e della conoscenza. Ad esempio, aiuta l'utente a generare link per la "navigazione", a eseguire ricerche semantiche, a disambiguare e definire esplicitamente il significato dei termini che l'utente usa nelle etichette per la classificazione, negli attributi dei documenti, nelle query di ricerca, e cosi` via. La conoscenza di base è completamente configurabile e puo` essere aggiornata dall'utente relativamente ad un task concreto.
SMatch [9.6] è uno strumento per il matching basato su schemi sviluppato all'Univeristà di Trento.
Prende due strutture ad albero (es. classificazioni) e restituisce le relazioni semantiche (es. equivalenze, sussunzione) tra i nodi dell'albero semanticamente corrispondenti.
Le relazioni sono determinate dall'analisi del significato (concetti, non etichette) che è codificato negli elementi e nelle strutture degli schemi/ delle ontologie. In particolare, le etichette dei nodi, scritte in linguaggio naturale, sono tradotte in formule proposizionali che codificano esplicitamente il loro significato.
Questo consente di codificare il problema del matching come un problema di soddisfacibilità nella logic proposizionale, che può essere efficacemente risolto utilizzando algoritmi di decisione (validi e completi) allo stato dell'arte. Sebbene sia stato principalmente creato per risolvere i problemi di eterogeneità semantica nell'integrazione dei dati/dell'informazione, SMatch puo` essere usato per altre applicazioni, ad esempio, nel ricercare utenti con gli stessi interessi, scoprire link per la "navigazione" tra le classificazioni degli utenti, e così via. Al momento, SMatch è una componente del sistema Sweb.

5. Descrizione generale delle attività

Il progetto è organizzato in cinque attività:

1. Estrazione di contenuti
2. Integrazione di contenuti
3. Presentazione di Contenuti
4. Piattaforma per la Gestione dei Contenuti e
5. Casi d'uso e diffusione.

Le prime tre attività coprono i principali interessi scientifici del progetto. Il loro obiettivo è quello di portare risultati nella ricerca e nuove soluzioni da integrare nella piattaforma. Ci aspettiamo che queste soluzioni siano applicabili in altri contesti oltre quello delle

Memorie Digitali. Inoltre, tali attività di ricerca hanno l'intento di creare opportunità di formazione, tramite l'impiego di studenti di dottorato nel progetto. La quarta attività, la Piattaforma Integrata, è dedicata all'integrazione delle soluzioni software sviluppate nell'ambito delle altre attività in un sistema ingegnerizzato. Infine diversi casi d'uso saranno costruiti sopra la piattaforma. Il progetto si articola in tre cicli di attività di un anno ciascuno. Durante il primo ciclo ci focalizzeremo sull'elaborazione di fonti istituzionali di dati, con l'estrazione dei dati di "background" ai quali gli utenti di LiveMemories avranno accesso; inoltre verrà sviluppata una prima versione della piattaforma integrando componenti esistenti. Durante il secondo ciclo, completeremo le prime funzionalità avanzate di visualizzazione e svilupperemo un primo caso d'uso di "Esempio" per illustrare le funzionalità della piattaforma e stabilire contatti con una comunità potenziale. Infine, nel terzo ciclo, realizzeremo un Caso d'Uso Pilota con questa comunità.

1. Estrazione dei contenuti. Ad oggi è possibile estrarre da testi informazioni su larga scala sia riguardo a entità (persone, luoghi, organizzazioni) sia riguardo a semplici relazioni tra entità (per es. l'affiliazione ad una organizzazione), ma con una precisione non molto soddisfacente. Il nostro intento è quello di migliorare significativamente lo stato dell'arte nell'estrazione di informazioni ed estendere l'attività al parlato e/o a documenti multilingui. La sfida principale che prevediamo è quella di sviluppare metodi che possano raggiungere un'alta precisione in presenza di dati molto eterogenei e potenzialmente molto sporchi; ci aspettiamo che il lavoro si concentri principalmente nell'area del data cleaning, della coreferenza intratestuale e dell'estrazione di relazioni. Tecniche consolidate di estrazione di contenuti da documenti visuali (immagini e video) saranno adattate allo scenario delle memorie digitali, in particolare riguardo ai requisiti richiesti per l'elaborazione, l'integrazione e l'accesso all'informazione.

2. Integrazione dei contenuti. L'obiettivo di questa attività è lo sviluppo di nuove tecniche per l'integrazione di grandi quantità di conoscenza, annotata temporalmente e possibilmente inconsistente estratta da una varietà di risorse, nella base di conoscenza che costituirà il nucleo di LiveMemories. Un obiettivo sarà lo sviluppo di nuovi metodi per la coreferenza intertestuale in grado di trattare un grande numero di dati e di informazioni temporali. Nuove forme per la rappresentazione di entità saranno necessarie: gli standard ad oggi adottati dalla comunità Web (come RDF e OWL) non garantiscono robustezza sufficiente per gestire e integrare in modo automatico i dati estratti da risorse non strutturate. La sfida di questa attività è la definizione e la sperimentazione di estensioni ai formalismi attuali, in grado di trattare fenomeni rilevanti quali il ruolo del contesto in cui l'informazione è calata, l'incertezza, l'incompletezza e la dipendenza dal tempo dell'informazione stessa. Un'ulteriore linea di ricerca che intendiamo seguire riguarda la convergenza tra gli approcci di inferenza basata su logica e la recente direzione, emersa nell'ambito della linguistica computazionale, basata sull'implicazione testuale.

3. Presentazione dei Contenuti. Un importante obiettivo per il progetto è la promozione di comunità che estrarranno informazioni relative ai loro temi di interesse, che le organizzeranno e le presenteranno ad altri (per es. in esibizioni virtuali). In questi termini il progetto ha una forte connotazione Web 2.0, migliorata dalla possibilità di usare la semantica e presentare il materiale tramite avanzate tecniche di visualizzazione, quali una presentazione adattiva basata sulla semantica, diverse tecniche di creazione automatica di riassunti (ad es. la creazione di testi riportanti come un individuo o una comunità si sono sviluppate nel tempo) e la traduzione automatica in lingue diverse dall'italiano. Alcune di

queste attività si baseranno su tecnologia esistente, mentre in altre l'avanzamento sarà reso possibile dalla ricerca portata avanti dal consorzio. Il mantenimento di viste multiple sulla conoscenza sarà un punto chiave, così come la ricerca sul matching semantico e sulle interfacce conversazionali. L'esperienza di Southampton nella visualizzazione multimediale sarà un punto chiave. La conoscenza acquisita dall'uso di molti strumenti software per gli studi sociali apparsi negli ultimi anni sarà preziosa per modellare le funzionalità offerte agli utenti e le interfacce.

4. Piattaforma per la gestione dei contenuti. Il cuore tecnologico del progetto è una biblioteca digitale multimediale per l'acquisizione, la gestione e l'integrazione di un grande numero di memorie collettive attive. Questa piattaforma permetterà una massiccia estrazione di contenuti e al contempo fornirà agli utenti le funzionalità di creazione e accesso a tali memorie, stabilendo dei collegamenti tra ogni tipo di informazione. Il risultato consisterà in un catalogo di entità spazio-temporalmente referenziate presenti nell'area selezionata per il caso d'uso e di una mappa delle relazioni esistenti tra queste entità. Sarà adottata un approccio centrato sulle entità, in cui è possibile cercare ed accedere alla conoscenza usando identificatori (per es. URI) per recuperare le entità; a loro volta i profili delle entità vengono costruiti raccogliendo da diverse fonti di conoscenza ciò che il Web ha da offrire su queste entità. Il progetto sfrutterà le tecnologie sviluppate per il progetto Ontotext (estrazione di conoscenza da testi), piattaforme GIS per mappe territoriali, comunità web e piattaforme per la condivisione di immagini.

5. Caso d'uso e diffusione. LiveMemories si concentra non tanto sulla vita dei singoli individui quanto sugli eventi della vita collettiva di una comunità. L'obiettivo è mettere a disposizione delle comunità una tecnologia innovativa per favorire la costruzione creativa di una memoria collettiva. Verrà reso disponibile al pubblico un portale Web per la raccolta, la gestione, l'integrazione e la fruizione di memorie collettive multimediali provenienti da diverse fonti (per es. associazioni, giornali e radio locali, ecc.), con il coinvolgimento anche di famiglie e singole persone. In un esperimento controllato, una comunità locale sarà supportata nella creazione delle proprie memorie collettive. Verranno organizzati eventi per coinvolgere la popolazione locale e per mostrare che gli sviluppi tecnologici, derivanti dall'idea di memorie come frammenti di contenuti digitali attivi, favoriscono nuove modalità di fare esperienza del nostro passato.

6. Originalità e rilevanza

LiveMemories produrrà contributi originali alle aree delle Tecnologie del Linguaggio, Gestione della Conoscenza, e Web Science, che renderanno il progetto di grande rilevanza scientifica. Gli argomenti di ricerca selezionati sono cruciali, e per essi sono stati individuati approcci innovativi che combinano metodi guidati dai dati con tecniche basate su conoscenza. I partner del consorzio sono esponenti riconosciuti della ricerca di punta nell'ambito della coreferenza intertestuale, dell'applicazione di metodi basati su kernel all'estrazione di relazioni, dell'implicazione testuale, della traduzione automatica statistica, del matching di ontologie e del ragionamento contestuale, e del trattamento e visualizzazione di informazione multimedia.

In primo luogo, riteniamo che i tempi siano maturi per affrontare su larga scala l'estrazione e l'integrazione di contenuti da fonti di informazione non strutturata. Mentre la comunità di ricerca si è focalizzata fino ad ora su benchmark di dimensioni ridotte (cfr. ACE) l'originalità del contributo del progetto LiveMemories consiste nell'applicare tale ricerca in uno scenario reale. L'opportunità, e l'ambizione, è di applicare tecnologie allo stato dell'arte ad un territorio di 500,000 abitanti come la Provincia di Trento. A nostra conoscenza, questo sarebbe il primo esperimento mai tentato di individuare e monitorare entità e fatti su una scala così vasta. L'originalità dell'approccio consiste nel fatto che il processo complessivo di popolamento dell'ontologia è un processo dinamico, in cui i fatti già riconosciuti vengono usati dinamicamente per fornire informazione agli algoritmi di apprendimento. Nella nostra visione questo obiettivo è parte di un piano a lungo termine per la realizzazione di una piattaforma permanente per la gestione della conoscenza del territorio trentino.

L'argomento specifico del progetto, le memorie digitali, è altamente innovativo e non ancora sfruttato sotto diversi aspetti. Mentre esistono iniziative rivolte alla raccolta di memorie digitali di specifici eventi come depositi statici, la visione originale di LiveMemories è quella di interpretare automaticamente tali memorie e collegarle tra loro. Queste "memorie vive" hanno un altissimo potenziale di applicazione per forme di fruizione originali, come ad esempio la ricostruzione di linee temporali per eventi, l'associazione di persone, organizzazioni, o luoghi a eventi, il recupero di fatti associati a specifici luoghi. Se consideriamo questo potenziale integrato in tecnologie basate sul web, ormai accessibile a larghi strati della popolazione e in cui si possono formare spontaneamente delle comunità, l'impatto sociale ed emotivo del progetto potrebbe cambiare drasticamente il modo in cui pensiamo le nostre memorie nell'era digitale.

Allo stesso tempo, la quantità di dati resi disponibili dalle tecnologie di estrazione di contenuti imporrà lo sviluppo di nuove teorie di ragionamento robusto in condizioni di incertezza dei dati. Sebbene questo non sia un argomento nuovo nell'ambito dell'intelligenza artificiale, i modelli sviluppati fino ad ora sono stati testati solo su scala ridotta, e molto poco si sa su come questi modelli possano comportarsi se devono essere gestiti milioni di dati.

L'integrazione di dati estratti da fonti istituzionali con dati di singoli cittadini e famiglie rappresenta un altro aspetto originale del progetto. A nostra conoscenza una tale possibilità non è offerta da nessuna delle piattaforme basate su web che sono oggi disponibili per la condivisione di contenuti. In questa nuova prospettiva ci aspettiamo che le comunità migrino dalla dimensione personale (famiglia, amici) ad una dimensione collettiva (un quartiere, una scuola, una città). La sfida a lungo termine (che va oltre gli scopi del progetto) consiste nell'indurre sia i fornitori di dati istituzionali sia associazioni e cittadini a percepire la raccolta e la conservazione delle memorie digitali come un'impresa collettiva.

7. Collegamento con programmi di ricerca nazionali e internazionali

- i2010 Digital Library Initiative. Le memorie digitali stanno emergendo all'interno di i2010 – una iniziativa EU il cui scopo è di rendere le tradizioni culturali europee più interessanti. L'area di applicazione riguarda la progettazione di nuove tecnologie Web per collezionare, gestire e rendere disponibili grandi databse di contenuto multimediale.
- Web Science Initiative. Nella visione di Web Science (cfr. il recente discorso alla House of Representatives degli USA da Tim Berners-Lee) lo sviluppo di metodi avanzati per l'integrazione della conoscenza e lo studio delle implicazioni sociali del Web sono centrali. L'attenzione è su strutture sociali dove le persone possono assegnare etichette condivise a documenti o altre sorgenti di informazione e poi recuperarli usando tali etichette. Entrambi questi aspetti saranno centrali in LiveMemories.
- Entity Disambiguation. Questa area è considerata strategica dal governo USA poiché è stata promossa da molteplici finanziamenti per iniziative come the Summer 2007 ELERFED CLSP workshop on Entità Disambiguation, coordinato dal gruppo di UniTN e la prossima campagna di valutazione di ACE su cross-doc coreference.
- RTE. La visibilità del challenge Recognizing Textual Entailment è anche mostrata dal fatto che la prossima edizione verrà organizzata dal NIST come parte della nuova conferenza Text Analysis Conferente, coordinata da FBK e Bar-Ilan.
- FIRB-Israel project "Intelligent Technologies for Cultural Tourism and Mobile Education" LiveMemories avvierà una collaborazione tra FBK, UniTN, Haifa e Bar Ilan, dove il gruppo UniTN applicherà metodi per il matching ontologico alla interfaccia dei musei, mentre FBK lavorerà sulle implicazioni testuali (textual entailment).
- OKKAM. Sfrutteremo la sinergia con il progetto OKKAM, un EU-IP coordinato da UNITN, con o scopo di OKKAM è la progettazione e implementazione di una struttura (worldwide) che supporti il riuso di identificatori globali per ogni entità del Web.

8. Attività formative di giovani ricercatori e tecnici

Come detto nella sezione 8, negli ultimi anni UniTN e FBK hanno fatto partire una serie di iniziative congiunte di training nelle aree dell'HLTI e della knowledge technology, a partire dal livello di Master fino al PhD, che provvedono un grounding rigoroso in queste tecnologie sia in prospettiva industriale che in prospettiva accademica. Uno dei punti di forza della proposta è che, sfruttando questo curriculum, sarà possibile fornire numerose opportunità di training applicativo ed accademico. Per quanto la realizzazione dei sistemi finali sarà gestita da ricercatori esperti, la ricerca vera e propria e numerosi sottoprogetti saranno affidati prevalentemente a PhD dei programmi internazionali DIT e CIMEC, a

studenti del Master HLTI, e a giovani tecnici che verranno così formati. (Ci si attende tra i 20-30 PhDs, 10-15 studenti di Master, e un numero simile di tecnici.) Comprendendo sia aspetti applicativi che di ricerca pura, e grazie al coinvolgimento di gran parte dell'industria HLTI locale –che già finanzia diverse borse di PhD e Master questi giovani studenti e tecnici avranno l'opportunità di acquisire esperienza con ambedue i tipi di attività.
Reciprocamente, il personale dell'industria locale potrà avere accesso alla ricerca avanzata in tempi molto brevi.
Stiamo anche considerando la possibilità di una scuola estiva in HLTI per attrarre e valutare potenziali studenti. Questa opportunità segue il successo della campagna di valutazione EVALITA, organizzata da FBK nel 2007, in cui per la prima volta un considerevole numero di sistemi per la lingua italiana sono stati valutati per cinque task condivisi (POS tagging, disambiguazione di senso, parsing, riconoscimento di entità nominate e riconoscimento di espressioni temporali). Riteniamo che una alternanza annuale della scuola estiva Evalita, in cui gli studenti potrebbero fare pratica sperimentale e preparare dati annotati, con il workshop Evalita, con la campagna di valutazione, avrebbe un impatti di rilievo sullo stato dell'arte delle tecnologie del linguaggio per la lingua italiana. Sebbene la scuola estiva Evalita avrebbe una organizzazione nazionale e internazionale, abbiamo previsto nel progetto un budget per aiutare il lancio di questa iniziativa sotto la spinta di LiveMemories.

9. Potenzialità di ricaduta sul contesto sociale e/o economico locale

LiveMemories si propone di avere un impatto sul contesto locale a diversi livelli:

- grandi quantità di memorie relative alla comunità locale verranno raccolte, digitalizzate e immagazzinate in modo da garantire la loro conservazione. La rete di musei storici locali sarà coinvolta in tutte le fasi di questo processo.
- la Piattaforma di Gestione dei Contenuti che sarà distribuita all'interno di Attività 4 rappresenterà un passo significativo nella direzione di un servizio permanente di monitoraggio dei fatti riguardanti il territorio provenienti da fonti di informazione multiple.

I gruppi a cui il progetto si rivolge coinvolgono pressoché chiunque in Trentino, con l'unico limite rappresentato dall'abilità che gli utenti avranno nell'immaginarsi e creare comunità. Gruppi utenti rilevanti sono:

- Cittadini del trentino, che saranno coinvolti in uno sforzo comune di raccolta e condivisione di memorie delle città. Il progetto rappresenterà un'opportunità unica di stimolare le comunità a partecipare ad un'impresa collettiva. Ci aspettiamo che la condivisione di un obiettivo comune permetterà di superare la distanza percepita tra istituzioni e popolazione.

- Turisti: la creazione di comunità di turisti renderà il Trentino più interessante, con la sua insolita combinazione di alta tecnologia e risorse naturali. Questo legherà i turisti ad una comunità più larga, aumentando il loro coinvolgimento nel territorio.
- Studenti e scuole, creando in tal modo un futuro per la nostra comunità e gettando le basi per un coinvolgimento attivo e proattivo di tutti i cittadini in un progetto che può durare per tutta la vita. Le scuole saranno coinvolte attivamente.
- Aziende del Trentino, con l'obiettivo di sfruttare il materiale multimediale prodotto dal progetto, attraverso lo sviluppo di specifiche applicazioni verticali in aree come il turismo, l'e-government e la sicurezza.

L'impatto del progetto sarà amplificato da un dibattito culturale, attraverso l'organizzazione di workshop scientifici, conferenze ed eventi aperti al pubblico. Il nostro scopo è quello di portare avanti un dibattito interdisciplinare su due livelli: (i) argomenti specifici tra quelli sviluppati nel progetto; (ii) l'impatto che LiveMemories ha su temi quali il ruolo delle tecnologie web nelle memorie digitali, l'impatto sociologico delle comunità, le nuove prospettive offerte dall'Iniziativa della Web Science. Abbiamo pianificato di organizzare una conferenza su questi temi con Tim Berners Lee, l'inventore del Web, e Wendy Hall, l'unica rappresentante nella commissione ERC con esperienza in informatica, entrambi professori a SOTON.
Infine, ci aspettiamo un impatto positivo sul contesto della ricerca del territorio, grazie alla collaborazione tra FBK e UniTN, in particolare nell'area HLT tra i gruppi di FBK, CIMEC e DIT.

10. Diritti di proprietà intellettuale preesistenti e loro impatto sulla valorizzazione dei risultati

Nel caso la proposta progettuale sia valutata positivamente, prima dell'inizio del progetto, il coordinatore (FBK) e i partners (UNITN e SOTON) sottoscriveranno un contratto in cui saranno definiti l'organizzazione del lavoro tra le parti, la struttura di gestione del progetto, diritti e doveri delle Parti, incluse responsabilità e diritti di proprietà.

TECNOLOGIE DI BASE E RISORSE
Gli IPRs per le tecnologie di base e le risorse usate nel progetto sono di proprietà dei partner. La maggior parte delle tecnologie già sviluppate (conoscenza pregressa) e tutta la conoscenza che verrà sviluppata nel progetto (conoscenza nuova) saranno messe a disposizione per scopi di ricerca al fine di massimizzarne la diffusione e l'impatto nella comunità scientifica. La conoscenza di background
disponibile include:

- SWEB: un'infrastruttura distribuita per la creazione di classificazioni di documenti, comunità, e cluster di luoghi ed eventi, e per la ricerca semantica al loro interno. Tutta la conoscenza nuova sviluppata usando SWEB sarà resa disponibile.
- TextPro (text processing di base per italiano e inglese), MultiWordNet (database lessicale per l'italiano) e I-CAB, un corpus italiano annotato da FBK, sono distribuiti per scopi di ricerca. È stato avviato il processo per rendere disponibile in

formato open source, attraverso SourceForge, il toolkit per la coreferenza ELKFED/BART.
- Il corpus di riferimenti anaforici per l'italiano VENEX, prodotto da UniTN, sarà a disposizione dei partner.
- MOSES (traduzione automatica), è distribuito secondo uno schema Open Source.
- Tecnologia di trascrizione automatica del parlato sviluppata presso FBK è stata concessa in licenza a Pervoice, che sarà coinvolta in subappalto per fornire servizi di trascrizione da notiziari e Consigli Provinciali.

TECHNOLOGIE DI SUPPORTO
LiveMemories dovrà integrare le tecnologie di base con piattaforme esistenti, per gestire luoghi geografici e comunità. Un candidato per i primi è Google Maps, le cui API sono già disponibili al pubblico, il che permette una loro facile integrazione nelle applicazioni. Per la gestione delle comunità, si potrebbero utilizzare le API di Flickr, una piattaforma sviluppata da Yahoo e già ampiamente usata; un eventuale accordo con Yahoo potrebbe prevedere un'instanziazione specifica di Flickr per LiveMemories.

DATI DI PROVIDER LOCALI
I dati usati nel progetto (es. notiziari e articoli) sono soggetti ai diritti di proprietà imposti dai provider.
Seguendo uno schema già applicato nel progetto Ontotext, intendiamo raggiungere accordi con essi affinché rendano disponibili i dati a scopo di ricerca durante il progetto.
Poiché le questioni di privacy legate al trattamento dei dati di soggetti non istituzionali potrebbero costituire una fonte di rischio per il progetto, intendiamo coinvolgere fin dall'inizio del progetto legali esperti nel campo; questi costi sono previsti nel budget.

11. Consorzio di Progetto: Schede soggetti partecipanti

11.1 FBK-irst

Descrizione delle competenze scientifico-tecnologiche

La Fondazione Bruno Kessler (FBK), nata il primo marzo 2007, è un ente privato senza fini di lucro con obiettivi di interesse pubblico che eredita la storia dell'Istituto Trentino di Cultura (ITC - fondato nel 1962 dalla Provincia Autonoma di Trento). Eccellenza scientifica, innovazione e trasferimento tecnologico sono gli obiettivi principali di FBK. Le attività di FBK - Ricerca Scientifica e Tecnologica vengono svolte all'interno di tre aree principali: Tecnologie dell'Informazione, Microsistemi e Fisica Applicata. Il personale di ricerca è formato da 80 persone assunte a tempo indeterminato e circa 50 persone con contratto temporaneo. Il budget ammonta a circa 19 milioni di Euro. La metà dei costi diretti è coperta da commesse industriali e da contratti nazionali ed europei. Ad oggi, FBK ha concluso più di 50 contratti europei.
All'interno dell'area di Tecnologie dell'Informazione, due unità di ricerca sono direttamente coinvolte nella proposta: l'unità di ricerca di Tecnologie del Linguaggio Umano (HLT) e l'unità di ricerca di Gestione dei Dati e della Conoscenza (DKM).
La prima unità è nata da tre linee di ricerca delle precedenti divisioni TCC e SSI dell'ITC-irst, mentre la seconda unità è nata da una linea di ricerca della precedente divisione SRA.

Le competenze dell'unità HLT si collocano principalmente nelle seguenti aree: trascrizione automatica del parlato, riconoscimento del parlato, estrazione di informazione linguistica da dati audio, elaborazione di informazione interlinguistica, traduzione automatica, interazione verbale in ambiente disturbato, estrazione di informazioni da testi (in particolare estrazione di entità e di relazioni tra entità), question answering, disambiguazione, acquisizione lessicale e sviluppo di risorse lessicali multilingue. L'unità HLT ha sviluppato tecnologia allo stato dell'arte in molti degli ambiti di ricerca in cui si colloca ed ha ottenuto ottimi risultati in molte campagne di valutazione internazionali, come DUC (creazione automatica di riassunti, miglior sistema per qualità linguistica nel 2005), SENSEVAL (disambiguazione, miglior sistema non supervisionato per il compito "all words" per l'inglese, secondo miglior sistema supervisionato per spagnolo, italiano e catalano), CLEF (question answering interlinguistico, dal 2004, miglior sistema italiano/italiano), TREC (question answering inglese, dal 2002, quarta posizione nel 2003, primo sistema europeo), PASCAL-RTE (riconoscimento di implicazione testuale, due partecipazioni), TERN (espressioni temporali, secondo sistema nel compito completo di riconoscimento e normalizzazione di espressioni temporali nel 2004), NIST-MT (traduzione automatica, dal 2003, quarto per arabo/inglese nel 2006), IWLST (traduzione di linguaggio parlato, miglior sistema nel 2005). Il gruppo che lavora sulla traduzione di testi e di parlato è attualmente coinvolto nel progetto open source più importante all'interno della comunità di traduzione automatica. La ricerca sul riconoscimento del parlato si colloca ai livelli più alti ed ha in varie occasioni raggiunto il mercato. La ricerca sull'estrazione di contenuti ha ottime pubblicazioni e risultati di valutazione, precisamente nei compiti di question answering e di estrazione di informazioni. Inoltre, alcuni ricercatori dell'unità svolgono ruoli chiave all'interno di molte iniziative internazionali incentrate su valutazione e creazione di benchmark (CLEF, PASCAL-RTE, IWLST, EVALITA). Infine, il lavoro sui modelli cognitivi, focalizzato sul cosiddetto affective computing sia su testi sia su parlato, ha portato come risultato pubblicazioni alle conferenze più importanti.
Le competenze dell'unità DKM comprendono: rappresentazione della conoscenza e ragionamento automatico, tecnologie di web semantico per la rappresentazione della conoscenza, acquisizione di conoscenza, matching semantico di ontologie eterogenee, logica per ragionamento distribuito, apprendimento automatico sistemi di raccomandazione.

Descrizione del gruppo di ricerca dedicato al progetto

Il gruppo FBK coinvolto nel progetto è formato da tre attori principali: le unità di ricerca HLT e DKM per l'attività di ricerca, e l'Ufficio Relazioni Territoriali per l'attività sullo showcase. La seguente descrizione del gruppo è focalizzata sulle attività direttamente connesse con il progetto.

UNITÀ DI RICERCA HLT
Tra tutte le attività di ricerca condotte dall'unità HLT, vengono descritte solo quelle direttamente coinvolte nel progetto.

Riconoscimento del parlato
L'attività verte sui seguenti argomenti. (1) Tecnologia di base: normalizzazione di caratteristiche acustiche basata su cluster non supervisionato; metodi di selezione dei dati per la modellizzazione del linguaggio; individuazione e riconoscimento di parole sconosciute attraverso unità inferiori alla parola; espansione e ripesatura di un grafo di

parola attraverso la morfologia; modellazione veloce di pronuncia per nuove lingue. (2) Analisi del contenuto del parlato: estrazione di caratteristiche prosodiche informative ed affidabili; metodi per valutare l'adeguatezza dei dialoghi rispetto a protocolli di servizio nonché la competenza linguistica degli studiosi. (3) Tecnologia multilingue: ad oggi la tecnologia sviluppata copre tre lingue, italiano, inglese e spagnolo.

Traduzione automatica
Questa attività comprende ricerca nell'area della traduzione sia di testi che di parlato. (1) Modellazione statistica: (i) integrazione di conoscenza linguistica nel cuore del motore statistico di traduzione; (ii) uso del contesto nella traduzione di documenti per migliorare la coerenza nella traduzione; (iii) metodi per gestire coppie di traduzione sottodocumentate al fine di superare il collo di bottiglia dei dati tipico degli approcci statistici. Alcuni risultati sono rilasciati come open-source (Moses, IrstLM). (2) Traduzione del parlato: (i) miglioramento di interfacce tra riconoscimento del parlato e traduzione per ridurre la propagazione; (ii) miglioramento della qualità dell'output della traduzione del parlato per avvicinarsi alle trascrizioni fatte a mano. (3) Acquisizione di conoscenza da dati: (i) adattamento di modelli di traduzione generici a contesti/dominii specifici per permettere la focalizzazione su argomenti specifici; (ii) acquisizione automatica di dati bilingui da dati comparabili; (iii) tecniche di selezione dei dati da grossi corpora al fine di abbassare la complessità e di permettere l'adattamento a dominii applicativi specifici; (iii) metodi di gestione delle parole nel documento di partenza che non sono state osservate nei dati di addestramento.

Elaborazione del contenuto
Questa attività copre la ricerca nelle aree di question answering e di acquisizione e integrazione di contenuti da documenti di testo. Le tecniche sviluppate sono applicate in casi d'uso concreti forniti dagli utenti coinvolti nei vari progetti (1) Question Answering: (i) sviluppo di un approccio basato su implicazione per domini di QA chiusi; (ii) partecipazione a campagne di valutazione di QA di dominio aperto (sia compiti monolingui che interlingui). Le attività orientate all'applicazione hanno lo scopo di sviluppare dimostratori di QA (con interfacce per dispositivi mobili e fissi) basati su un'architettura distribuita di servizi web. (2) Acquisizione di conoscenza: estrazione di locuzioni chiave, estrazione di entità e relazioni tra entità, apprendimento e popolamento di ontologie, integrazione di tecniche testuali all'interno di un paradigma per l'acquisizione di conoscenza multimediale e intermediale, acquisizione automatica di sequenze di inferenza testuale. In questo compito sono usate sia tecniche basate su regole sia basate su apprendimento automatico. (3) Integrazione di contenuti: (i) coreferenza intra- e interdocumento tra nomi di entità (persone, organizzazioni, luoghi); (ii) coreferenza di relazioni nel tempo. L'output del processo di integrazione è usato per implementare sistemi di accesso all'informazione innovativi in linea con il web semantico.

Infrastruttura
Questa attività fornisce supporto tecnologico e servizi di alto livello al fine di ottimizzare le attività dell'unità HLT. (1) Infrastruttura tecnologica. Include la gestione di cluster di computer ad alte prestazioni, l'installazione e la gestione di specifici pacchetti di software (ad es. processori linguistici), l'immagazzinamento e il recupero di grandi quantità di dati (ad es. dati acustici) così come la definizione del loro formato e la documentazione, l'infrastruttura per sistemi di dimostrazione basati sul web. (2) Risorse linguistiche. Svolge due tipi di attività: (i) sviluppo e mantenimento di risorse multilingui, scritte e di parlato, (ii) disseminazione e mantenimento della rete di relazioni. La prima attività

include progettazione e raccolta dei dati, definizione di schemi di annotazione, annotazione automatica, creazione di dati di addestramento e di valutazione, mantenimento e gestione delle risorse. La seconda attività è focalizzata sul mantenimento di relazioni con altre istituzioni, distribuzione delle risorse, disseminazione dei risultati e organizzazione di eventi e di campagne di valutazione.

UNITÀ DI RICERCA DKM
L'attività di ricerca consiste nello sviluppare modelli logicamente e computazionalmente sostenibili per supportare la creazione di conoscenza da dati, l'integrazione della conoscenza e la creazione di servizi di ragionamento. (1) Elicitazione della conoscenza. La ricerca si concentra sulla estrazione di conoscenza da dati, come dati testuali preprocessati (ad es. documenti testuali e multimediali annotati), dati con una semantica leggera (ad es. tassonomie di persone e wiky), dati strutturati (ad es. database e file XML). Il risultato del processo di elicitazione della conoscenza è una teoria logica processabile automaticamente, chiamata modulo di conoscenza. (2) Integrazione della conoscenza. Vengono presi in esame tre fattori principali: eterogeneità (dati immagazzinati in diversi moduli di conoscenza possono avere una semantica diversa), autonomia (la gestione di moduli di conoscenza può non essere centralizzata), scalabilità (la dimensione e il numero dei moduli possono essere estremamente alti. (3) sviluppo di servizi di ragionamento logico ed euristico, loro applicazione a supporto dell'estrazione e integrazione automatica di contenuto, di composizione di servizi di web semantico, di analisi di procedure mediche.

UFFICIO RELAZIONI TERRITORIALI
L'ufficio ha la funzione di tenere tutte le relazioni utili con i numerosi interlocutori di FBK. La molteplicità delle specializzazioni della ricerca interna a FBK - tecnologica, scientifica e umanistica - grazie ai suoi 40 anni di storia, ha creato una vasta rete di relazioni su piani diversi. Le Relazioni Territoriali rispondono all'obiettivo di rendere le relazioni storiche con altri soggetti della ricerca e della formazione, con il mondo delle imprese e con le istituzioni pubbliche, concrete e positive. Ricerca di fondi, monitoraggio delle aspettative e dei bisogni dei vari interlocutori, disseminazione dei risultati di ricerca e una più ampia funzione di consulenza al territorio da parte delle competenze interne a FBK sono gli obiettivi specifici dell'ufficio. In particolare per i progetti di ricerca, le Relazioni Territoriali cercano il coinvolgimento attivo di quanti, soggetti istituzionali, pubblici e privati, possono trarre benefici dalla collaborazione in ordine alla crescita della cultura dell'innovazione, alla maggiore vicinanza con gli aspetti più strategici della ricerca scientifica internazionale, alla conoscenza del potenziale di sviluppo delle applicazioni possibili.

Collegamento con programmi di ricerca interni e posizionamento rispetto alla strategia interna

La ricerca avanzata e la tecnologia prodotta fina ad ora saranno sfruttate da FBK per far fronte alle nuove sfide di ricerca presentate dal progetto. I tre componenti del gruppo FBK si focalizzeranno su attività diverse in base ai loro specifici programmi di ricerca. L'unità HLT sarà responsabile dell'attività sull'estrazione di contenuti da fonti multimediali, lavorerà sulla coreferenza e sull'annotazione geo-temporale (attività di integrazione dei contenuti) e sarà responsabile della creazione della piattaforma di LiveMemories. L'unità DKM sarà responsabile dell'attività sull'integrazione dei contenuti e l'Ufficio Relazioni Territoriali lavorerà allo showcase.

Per quanto riguarda le strategie di ricerca, nel corso del progetto FBK aprirà nuove attività di ricerca nei campi delle tecnologie del linguaggio e del web semantico. Più precisamente sarà attivata una nuova linea di ricerca su integrazione di contenuti guidata dai dati, dove tecniche statistiche già applicate con successo nell'ambito delle tecnologie del linguaggio (ad es. riconoscimento di entità nominate) saranno sperimentate su un numero di fenomeni di più alto livello.
FBK manterrà anche la consolidata rete di collaborazioni strategiche che coinvolgono soggetti chiave nel territorio, tra cui l'Università di Trento (Master in HLTI, Dottorato in ICT), l'Università di Bolzano (Master in LCT);UniTN-CIMEC (ricerca congiunta su linguaggio e cognizione, pianificazione di un corso in Informatica Cognitiva), CELCT (organizzazione di campagne di valutazione) e PerVoice (spin-off HLT di tecnologia per il riconoscimento del parlato).
Oltre a contribuire dal punto di vista della ricerca, FBK avrà anche la responsabilità della coordinazione della Piattaforma di Integrazione dei Contenuti, un impegno congiunto con UniTN per costruire una biblioteca digitale multimodale per l'acquisizione e l'integrazione di contenuti. Questa attività sfrutterà e valorizzerà le competenze del servizio di infrastruttura dell'unità HLT.

11.2 Università di Trento

Descrizione delle competenze scientifico-tecnologiche

Dall'istituzione, nel 2001, delle graduatorie nazionali CENSIS, l'Università di Trento si classifica sempre tra le migliori università italiane, ed occupa il primo posto in molte aree scientifiche e tecnologiche.

DIT
Sebbene sia stato istituito soltanto nel 2002, il Department of Information and Communication Technology (DIT) ha un livello di produzione scientifica molto alto e grande capacità di attrarre fondi R&D dall'industria, dall'amministrazione locale e dall'EU. Il DIT ha partecipato a 22 progetti del 6º Programma Quadro, nonostante sia diventato un'entità indipendente solo durante la seconda parte del Programma. Tra il 2004 e il 2007, il DIT ha ottenuto più di 16M di Euro in fondi di ricerca, dall'industria o per l'educazione, mentre il supporto dall'università nello stesso periodo è stato di 6M di Euro. Il DIT ha sviluppato percorsi di studio molto efficaci che hanno ottenuto grande visibilità. La scuola di dottorato internazionale in ICT ha avuto un successo enorme; al momento, il programma coinvolge 155 studenti. Il DIT ha 74 membri tra il personale docente, 24 membri nello staff di ricerca, 14 post-doc, 10 tecnici e 19 impiegati amministrativi. La rappresentazione della conoscenza e il ragionamento sono sempre stati tra i punti di forza del DIT, e Web Science e Human Language Technology and Interfaces sono aree in forte espansione.

CIMeC
Il Centro Interdipartimentale Mente e Cervello (CIMeC) è stato di recente costituito per portare il Trentino all'avanguardia nella ricerca interdisciplinare alla frontiera tra neuroscienze, psicologia, informatica e fisica, e si sta rapidamente affermando come uno dei centri leader a livello europeo in questo settore; la scuola di dottorato internazionale in Cognitive and Brain Sciences sta già attirando una notevole quantità di studenti di alto livello dall'Italia e dall'estero. Il Language, Interaction and Computation Lab è uno dei

quattro laboratori in cui sono organizzate le attività del CIMeC. Include 5 docenti, 2 post-doc e 7 dottorandi. La ricerca del laboratorio si concentra sull'acquisizione di conoscenze di senso comune da vaste quantità di dati e sulle interfacce adattive.

DSSR
Il Department of Sociology and Social Research si dedica allo studio di macro-aree quali la struttura
sociale; diseguaglianze e azioni collettive; norme sociali e valori politico-etici; politiche sociali ed istituzioni europee; culture e cambiamento organizzativo; teoria sociale. Il dipartimento è fortemente impegnato in progetti di ricerca e attività nazionali e internazionali. L'unità di ricerca su Communication, Organizational Learning and Aesthetics (www.unitn/rucola) coinvolge studiosi e ricercatori che dal 1993 collaborano sulla base di interessi professionali comuni a studi su organizzazione e sistemi informativi, prediligendo metodi di ricerca qualitativi e interpretativi. L'unità include docenti, ricercatori indipendenti e dottorandi.

Descrizione del gruppo di ricerca dedicato al progetto

I gruppi UniTN coinvolti nella proposta sono conosciuti a livello internazionale in aree centrali alla riuscita del progetto, quali Web Science, Human Language Technology and Interfaces e Knolwedge Management and Reasoning.
Web Science è una nuova area in rapida espansione presso UniTN, dove le ricerche rilevanti si sono
concentrate su:

- Sviluppo di sistemi web-based multi-livello ed efficienti di tipo sia tradizionale che Web 2.0/Semantic Web, basati sull'utilizzo di tecnologia Web Service (per es., progetto ELEAF);
- Sviluppo di linguaggi per la rappresentazione della conoscenza sul Web. In collaborazione con l'IRST è stato sviluppato C-OWL, un linguaggio per l'allineamento di multiple ontologie scritte in OWL.

Il DIT coordina l'iniziativa LiquidPub (http://project.liquidpub.org/), che si propone di connettere e
formalizzare gli aspetti sociali e tecnici del paradigma di pubblicazione di materiali scientifici su grandi reti, quali il web. Al momento l'iniziativa si focalizza sullo sviluppo e sulla promozione di un nuovo paradigma per la produzione, disseminazione, valutazione ed uso della conoscenza scientifica. Questo paradigma, reso possibile dall'avvento del Web e da progressi ICT, propone la nozione di "Liquid Publications", ovvero contributi scientifici collaborativi, componibili ed in evoluzione. L'idea si basa in parte su paralleli tra artefatti della conoscenza e software, e dunque su quanto si è imparato dallo sviluppo (agile, collaborativo, open source) di software. Essa prende inoltre spunto dall'esperienza Web 2.0, quanto a valutazione collaborativi dei prodotti della conoscenza. Per informazioni ulteriori, si veda il sito http://project.liquidpub.org/.
Quella dell'Human Language Technology and Interfaces è stata una delle principali aree di crescita di UniTN, dove vengono sfruttate sinergie con l'IRST, che verranno ulteriormente rafforzate dalla proposta in corso. Il primo obiettivo in quest'area è stato quello di preparare corsi di studio internazionali indirizzati a chi sia orientato alla carriera accademica così come a chi sia più orientato al settore privato. Accanto al dottorato internazionale in ICT presso il DIT, di grande successo, caratterizzato dall'enfasi

tecnologica e dalla forte presenza IRST, negli ultimi due anni è stato sviluppato un percorso di dottorato internazionale presso il CIMeC a Rovereto, che si focalizza sugli aspetti cognitivi, ha forti componenti linguistiche e di studio dell'interazione ed enfatizzerà anche lo studio delle interfacce. Il nuovo Master HLT and Interfaces (in collaborazione con l'IRST) ha inizio nell'anno accademico corrente e si prefigge lo scopo di formare personale per l'industria locale, in continua crescita, e preparare ai dottorati gli studenti più inclini alla ricerca. Ad oggi, queste iniziative hanno portato all'assunzione di cinque docenti ed hanno avuto un'ottima ricezione, sia in termini di numero e qualità degli studenti interessati che in termini di supporto da ditte locali, quali Cogito e PerVoice, che sostengono i programmi con borse di studio. Il presente progetto porterà anche a cementare i rapporti con queste ditte. Ulteriori iniziative per il rafforzamento di quest'area sono in corso.
I gruppi DIT e CIMeC hanno un'ottima reputazione nel HLT, in particolare per quanto concerne
l'elaborazione dei testi e del parlato, la traduzione automatica su base statistica e le interfacce,
concentrandosi tra l'altro su:

- Creazione ed utilizzo di strumenti d'elaborazione del linguaggio naturale per tokenizzazione, annotazione morfostinattica, lemmatizzazione
- Text mining: estrazione di named entities, relazioni ed ontologie, annotazione di ruoli semantici (con partecipazione in competizioni internazionali quali CoNLL)
- Disambiguazione di entità con applicazioni HLT (membri di UniTN hanno co-organizzato il 2007 John Hopkins workshop su corerefenza e disambiguazione di entità, di grande successo)
- Creazione, ripulitura ed annotazione di corpora, dalla annotazione a mano d'alta qualità di corpora di medie dimensioni (per es., GNOME, ARRAU, LUNA) all'annotazione automatica di corpora di grandissime dimensioni (per es., partecipando all'iniziativa internazionale WaCky)
- Sviluppo di metodi di machine learning applicati all'HLT, e sviluppo di strumenti corrispondenti (per es., per active learning e metodi kernel)
- Riconoscimento del parlato in compiti con vocabolario molto esteso: i sistemi UniTN si sono classificati tra i primi tre in competizioni internazionali (per es., DARPA ATIS e EARS)
- Comprensione del linguaggio parlato ed interfaccia a basi di dati (per es., ATIS), con un sistema che si è classificato primo nella valutazione internazionale DARPA; UniTN partecipa inoltre al progetto LUNA (finanziato dall'EC) sulla comprensione del parlato in contesto multilingue
- Traduzione automatica su base statistica con corpora bilingui di grandi dimensioni (per es.,
  francese-inglese e arabo-inglese), utilizzando automi a stati finiti (progetto DARPA TIDES)
- Interfacce uomo-macchina che vanno da sistemi conversazionali multilingui per dialogo ad iniziativa mista e input parlato spontaneo (per es., il progetto “How May I Help You?”) a interfacce web; il team UniTN ha vinto un fondo Marie Curie Research Excellence per sviluppare le interfacce uomo-macchina della prossima generazione Knowledge Management, il gruppo DIT che si occupa di conoscenza e ragionamento, è noto a livello internazionale per studi nelle seguenti aree:
- Ragionamento contestuale e prospettive molteplici sulla conoscenza

- Ontologie "leggere" che permettano di tradurre classificazioni standard, quali Dmoz, in classificazioni formali, ovvero grafi etichettati in un linguaggio concettuale proposizionale. In questo modo compiti essenziali, quali la classificazione di documenti e query answering, si riduce ad un caso di ragionamento basato su sussunzioni
- Il Matching Semantico come operazione che opera su due grafi (per es., classifcazioni sul web, cataloghi commerciali, ontologie) e produce una proiezione tra i nodi che hanno una corrispondenza semantica.

L'apprioccio permette di trasformare il problema di matching in un problema di validità proposizionale, risolubile usando risolutori proposizionali completi e coerenti allo stato dell'arte
Il lavoro dell'unità si concentra anche su:

- ragionamento automatico effciente con logiche modali e temporali e procedure di decisione SAT
- gestione peer-2-peer di dati e conoscenza
- gestione della conoscenza centrata sulle entità
- logiche per la rappresentazione della conoscenza

Communication, Organizational Learning and Aesthetics (RUCOLA) è un'unità nota internazionalmente per lo studio della costruzione sociale della tecnologia, l'esplorazione delle pratiche organizzative, gli aspetti collettivi, sociali, affettivi e non puramente cognitivi di conoscenza e apprendimento e una forte enfasi sulla relazione tra gli aspetti linguistici, simbolici materiali ed emozionali dei processi organizzativi.

Collegamento con programmi di ricerca interni e posizionamento rispetto alla strategia interna

Il passaggio da un web personale a uno comune, esplorato in questo progetto, crea opportunità che non sono sfruttate dai meto i per condividere la conoscenza implementati da Flickr, Facebook, ecc. I
documenti condivisi – testi e immagini – contengono ricche informazioni che non sono oggi usate per migliorare la navigazione, né per fornire agli utenti nuove prospettive, né per estrarre le conoscenze tipiche del senso comune cruciali in applicazioni HLT.
La ricerca sulla gestione della conoscenza ha rinunciato a schemi globali pre-definiti e si concentra oggi su classificazioni definite dall'utente. Questo richiede lo sviluppo di metodi per: (i) costruire ontologie da tali classificazioni tramite l'elaborazione del linguaggio naturale; (ii) ragionare su ontologie definite dall'utente per automatizzare compiti quali classificazione dei contenuti, ricerca semantica, rilevazione di incosistenze;
(iii) matching delle ontologie per trovare corrispondenze tra ontologie di utenti distinti; (iv) analizzare e sfruttare tali corrispondenze.
Le nostre ricerche in HLT si focalizzano su metodi per (i) estrarre conoscenza – in particolare, descrizioni di entità basate su relazioni – da grandi basi di dati; (ii) organizzare tale conoscenza, per es. attraverso disambiguazione delle entità; e (iii) passare dall'estrazione di conoscenza da testo e parlato all'integrazione con immagini e video (l'interazione con Southampton sarà di grandissima utilità).
Nello studio delle interfacce, il progetto si collega al nostro lavoro su medodi di (i) organizzazione, accesso (per es., via ricerca semantica) e presentazione adattiva (per es., attraverso narrazioni) della conoscenza – di nuovo, un'area in cui l'esperienza di

Southampton ci sarà di grande aiuto; e (ii) visualizzazione e gestione di ontologie (per es., per popolarle o modificarle). La sfida, qui, è quella di rendere il compito naturale per l'utente, riducendo la pesante curva d'apprendimento legata alle ontologie.

11.3 Università di Southampton

Descrizione delle competenze scientifico-tecnologiche

School of Electronics & Computer Science (ECS)
ECS è nel Regno Unito l'istituto leader nel settore, essendo valutato top 5 per la ricerca negli ultimi due Research Assessment Exercises, e avendo ottenuto il grado best 5 dal Higher Education Funding Council per l'Inghilterra nel 2003. Ha una considerevole presenza in molte aree attinenti; ha organizzato la conferenza WWW-2006, è stato il primo istituto accademico nel mondo ad adottare un mandato self-archiving e ha creato il primo e il più usato software di archiviazione, EPrints. Precedente Direttore dell'istituto è stata Wendy Hall. Il gruppo Intelligenza, Agenti, Multimedia è un gruppo interdisciplinare incentrato su progettazione e applicazione di sistemi per l'elaborazione di informazione e di conoscenza complessa, con esperienza in aree che comprendono grid computing, sistemi P2P, reti sensoriali, Web Semantico (WS) e ambienti di calcolo pervasivi. Tim Berners-Lee, direttore del W3C, è professore all'interno del gruppo. Nel 2004-5, ECS ha ricevuto un finanziamento per la ricerca di £22m e ha pubblicato 642 articoli.

Web Science
Il Web Science Research Initiative (WSRI) è uno sforzo congiunto tra CSAIL di MIT e ECS, con quattro direttori: Tim Berners-Lee, Wendy Hall, Nigel Shadbolt, professore di IA a Southampton e direttore di AKT (2000-2007), e Daniel Weitzner, Technology and Society Domain leader del W3C e principal research scientist al MIT. James Hendler, professore di informatica, è direttore associato.
WSRI è Web Science all'avanguardia che studia sistemi informativi decentralizzati. Web Science implica creare nuovi protocolli di infrastruttura e capire la società che li usa, per supportare il riutilizzo dell'informazione in contesti nuovi. È focalizzato sulla decentralizzazione per evitare colli di bottiglia sociali e tecnici, e integra potenti tecniche scientifiche e matematiche provenienti da varie discipline per considerare proprietà microscopiche, fenomeni macroscopici e loro relazioni. La ricerca di IAM in quest'area si è focalizzata sulla capacità di mantenere l'informazione separata dai collegamenti (ipermedia aperto) in modo tale da poter essere personalizzata per ciascun utente (ipermedia adattivo) usando sofisticati modelli del significato di documenti, dati e loro interconnessioni (WS). Il gruppo è all'avanguardia nello sviluppo del WS per spostare il focus del Web dai documenti ai dati.

Memories for Life
Memories for Life (M4L) è una rete fondata da EPSRC (2004-7) per riunire un gruppo di accademici nel tentativo di comprendere il funzionamento della memoria e per sviluppare tecnologie per migliorarla, ora che la memoria umana è integrata da una crescente quantità di informazioni digitali personali; e-mail, fotografie, telefonate via Internet, località GPS e log di visioni televisive. M4L riunisce psicologi, neuroscienziati, sociologi e informatici per studiare modi efficaci di usare e gestire sia la memoria umana che quella computerizzata.

Multimedia Imaging
IAM ha una forte reputazione in diversi ambiti della ricerca multimediale, tra cui l'analisi del contenuto di multimedia, come immagini, video e realtà aumentata, per permettere di effettuare in maniera più efficace ricerca, navigazione e recupero basati sul contenuto, e fornire un accesso diretto alla semantica. Studia inoltre architetture per sistemi multimediali per combinare l'estrazione automatica, la rappresentazione e la manipolazione di contenuto semantico con più tradizionali strumenti di gestione dell'informazione in ambienti distribuiti. Un altro focus è su strumenti e tecniche per facilitare l'estrazione di informazione da database visuali e l'interpretazione e la presentazione di tale informazione con ipermedia. L'attività corrente include recupero di immagini ad alta risoluzione per collezioni di gallerie d'arte, gestione di metadati continui per flussi di media continui, reperimento basato su contenuto di immagini 2D/3D e combinazione di immagini e scansioni 3D per opere d'arte.

Descrizione del gruppo di ricerca dedicato al progetto

IAM è uno dei dieci gruppi di ricerca a Southampton che si occupano di Elettronica e Informatica. La ricerca di IAM è focalizzata sull'elaborazione di informazione e conoscenza complesse, con vari temi, tra cui agenti, grid computing e tecnologie della conoscenza. Il gruppo su LiveMemories ha esperienza in tre aree particolari: tecnologie della conoscenza, multimedia e Web Science (la scienza dei sistemi di informazione decentralizzati).

Tecnologie della conoscenza
La tecnologia della conoscenza è un tema fondamentale per IAM, che è incentrato su tecnologie,
formalismi e protocolli del WS. IAM è stato coordinatore del progetto Advanced Knowledge Technologies (AKT). AKT concettualizzava il ciclo della conoscenza come una serie di sei stadi: acquisizione della conoscenza, modellizzazione, recupero, riutilizzo, pubblicazione e mantenimento. Il gruppo ha sviluppato ed esteso tecnologie e standard per fornire metodi e servizi integrati per questi compiti.
L'approccio di AKT usa tecnologie del WS come mezzo di interrogazione di risorse. Questa ricerca ha portato a diversi strumenti e servizi intelligenti per creare e mantenere contenuti sul Web, e per contribuire alla creazione di documenti, insiemi di dati e altre risorse basate su conoscenza.
Un altra importante attività riguarda la creazione di contenuti annotati nella forma richiesta dal SW, usando metodi di Natural Language Processing (NLP), estrazione di conoscenza mediata dall'ontologia, annotazione automatica e semi-automatica e strumenti per aiutare l'annotazione (anche di tipo multimediale). Southhampton ha collaborato da vicino con ricercatori di NLP di altre università e riconosce l'importanza dell'uso di tecniche di NLP sul Web e per gestire grandi repository di conoscenza. Nell'ambito di queste collaborazioni sono stati sviluppati strumenti per l'analisi di testi che fanno uso di ontologie. La grande quantità di contenuti disponibili sul Web ci ha consentito di andare oltre l'annotazione incentrata sull'utente, verso l'estrazione non supervisionata e parzialmente supervisionata.
AKT ha sviluppato, sia per il Web che per repository di informazione su larga scala, importanti parti di infrastrutture a tutti i livelli (da tecnologie per immagazzinare dati a interfacce), tra cui metodi e strumenti per la ricerca e la navigazione e strumenti per cercare contenuti e per dare senso ai contenuti trovati. I metodi sviluppati sono stati e saranno applicati anche ad altri repository di conoscenza, tra cui archivi multimediali,

facilitando il riutilizzo di basi di conoscenza. Inoltre, questa infrastruttura è stata sfruttata in una ditta spinoff che si occupa di sicurezza dei dati e della privacy dei suoi proprietari, un aspetto chiave in un repository pubblico di memorie.
Un particolare punto di forza èlo sviluppo di strumenti, metodi e tecniche per lavorare con le ontologie – spesso considerate un potenziale collo di bottiglia per lo sviluppo del WS. Il lavoro di AKT sulle ontologie include metodi per la valuazione, il mapping, il merging, il mantenimento, la modularizzazione e la frammentazione.
Un altro importante filo conduttore della ricerca di AKT riguarda l'utilizzo di informazione del SW per indurre relazioni umane, istituzionali e organizzazionali. Sono stati sviluppati degli strumenti per identificare comunità di invidui, per localizzare determinate competenze e poi associare individui e competenze. IAM include anche ricercatori che hanno studiato problematiche relative all'affidabilità e provenienza dell'informazione, problematiche che stanno in primo piano nel caso di repository distribuiti di grandi dimensioni.

Multimedia
Esistono sempre piu' sistemi per immagazzinare, recuperare e processare informazione digitale multimediale, poiché gli strumenti per catturare immagini digitali, audio digitali e video digitali diventano meno costosi e piu' facili da usare. Lo sviluppo di strumenti versatili per gestire questo materiale in modo efficace offre molte sfide e opportunità di ricerca. Grazie all'elaborazione in tempo reale è possibile dare un supporto diretto a persone che interagiscono con il loro ambiente, in modo che le immagini, i video e gli audio possono essere usati per collegare il mondo digitale e quello fisico.
La ricerca di IAM include l'analisi del contenuto su un ampio spettro di media che includono immagini, video, audio e realtà aumentata. Mira a costruire sistemi per permettere di effettuare in maniera piu' efficace ricerca, navigazione e recupero basati sul contenuto. Si studiano strumenti piu' intelligenti per dare accesso diretto alla semantica dei media. Questo richiede di risolvere molti problemi nei seguenti campi: computer vision, comprensione audio (parlato e musica) e visualizzazione in 3-D. Le attività attuali del gruppo includono sistemi per il recupero di immagini ad alta risoluzione per collezioni di gallerie d'arte, gestione di metadati continui per flussi di media continui, recupero basato sul contenuto di immagini 2D/3D e combinazione di immagini e scansioni 3D per opere d'arte.
Combinando estrazione automatica, rappresentazione e manipolazione di contenuti semantici con strumenti piu' tradizionali per la gestione di informazioni in ambienti distribuiti richiede nuove architetture di sistema multimediali, e queste sono a loro volta studiate. Lo sviluppo di contenuti cross-mediali integrati e di sistemi per il recupero basato sui concetti di multimedia distribuiti continua a costituire un interessante obiettivo di ricerca a lungo termine.
Tecnologie della conoscenza e multimedia sono stati combinati nel progetto AKT con lo sviluppo di Photocopain, un sistema per la descrizione e archiviazione di esperienze personali, un task in cui gli utenti saltuari non sono molto propensi a fare lo sforzo necessario per fare le annotazioni richieste per organizzare le loro collezioni. Photocopain è un sistema semi-automatico per l'annotazione di immagini che combinando informazioni provenienti da diverse fonti, crea delle annotazioni che poi l'utente puo' ampliare o migliorare.

Web Science

Il Web è un esempio di un tipo di struttura di informazione che sta diventando sempre piu' comune, ovvero una miniera di informazioni decentralizzata e distribuita che fornisce agli utenti un mezzo per accedere a un numero sempre crescente di fonti di informazioni diverse (che vanno da archivi istituzionali a reti sensoriali utilizzate nel mondo reale). Tuttavia, usare queste informazioni per prendere decisioni presenta una serie di problemi. Le informazioni possono essere incomplete, non sicure o contraddittorie. Possono derivare da fonti possedute da persone con interessi diversi, e possono incorporare un gran numero media diversi. Risulta quindi evidente la necessità di avere sistemi che siano in grado di interagire con queste diverse fonti di informazioni non solo per raccogliere informazioni attinenti, ma anche per riformularle e ragionare su di esse.
Il Web è definito da poche semplici regole che danno vita a strutture molto complesse. Web Science mira ad analizzare le relazioni tra le regole semplici e il comportamento complesso, per identificare tendenze che possano minacciare o frammentare il Web, e per contribuire alla ricerca necessaria per assicurarne un continuo sviluppo. La ricerca di IAM si è focalizzata sulla capacità di tenere separati le informazioni e i collegamenti (ipermedia aperto) cosi da poter essere personalizzati per ciascun utente (ipermedia adattivo) usando sofisticati modelli del significato di documenti, dati e loro interconnessioni (il Web Semantico). Molti sistemi costruiti dal gruppo sono stati usati per l'ingegneria, l'editoria e l'istruzione e hanno contribuito a determinare gli attuali standard del Web. Il gruppo è sempre stato all'avanguardia nello sviluppo del Web Semantico come mezzo per spostare il focus del Web dai documenti ai dati.

Collegamento con programmi di ricerca interni e posizionamento rispetto alla strategia interna

La ricerca condotta in Livememories è complementare a diversi temi di ricerca condotti da IAM, tra cui tecnologie della conoscenza, estrazione di contenuti con tecnologie simili del SW e Memories for Life.
Fornendo un esempio di un grande sistema di informazioni decentralizzato, con contenuti forniti in modo distribuito, il progetto rappresenta anche un'opportunità per testare le attività di Web Science di IAM.
L'uso di tecnologie della conoscenza in un contesto multimediale (Photocopain, descritto sopra) ha costituito un'unione di tecnologie di successo, ma questo particolare sistema ha rappresentato solo una piccola parte della collaborazione AKT e quindi un primo passo. L'integrazione tra tecnologie per la conoscenza e multimedia dovrà essere esplorata più nel dettaglio, per comprendere meglio le possibilità di altre informazioni che sono a disposizione (quali ontologie, folksonomie, annotazioni di risorse correlate, etc.). Inoltre tali informazioni potrebbero essere usate in maniera più integrata. In questo senso il progetto aiuterebbe a creare sinergie all'interno del gruppo tra ricerca multimediale, tecnologie per la conoscenza, studio delle dinamiche di repository di conoscenza su larga scala decentralizzati. Rappresenterebbe inoltre una base sulla quale testare tecniche di analisi di immagini piu' dirette, usate a IAM, tra cui tecniche di machine learning e feature detector.
Nell'area Memories for Life, all'interno della quale IAM ha organizzato una rete EPSRC, verranno messe in evidenza le capacità operative di IAM. Poiché la raccolta di informazioni sta diventando sempre meno costosa, navigare grandi banche dati diventerà un bisogno primario non solo in contesti di business ma anche nella vita quotidiana e questo è un ambito di ricerca nel quale IAM auspica di mantenere la sua attuale posizione prominente.