PROGRAMMA DI RICERCA - MODELLO A
Anno 2004 - prot. 2004132117

1.1 Programma di Ricerca di tipo

Interuniversitario

Area scientifico disciplinare *Scienze economiche e statistiche (45%)*
Area scientifico disciplinare *Ingegneria industriale e dell'informazione (40%)*
Area scientifico disciplinare *Scienze matematiche e informatiche (15%)*

1.2 Titolo del Programma di Ricerca

Testo italiano
Metodologie di data mining per le applicazioni di e-business

Testo inglese
Data mining methods for e-business applications

1.3 Abstract del Programma di Ricerca

Testo italiano
Il progetto di ricerca ha come tema principale lo studio, lo sviluppo e l'implementazione delle metodologie di data mining più opportune per le applicazioni di e-business. Ci si occuperà prevalentemente degli aspetti metodologici ed applicativi inerenti l'analisi di dati provenienti dall'interazione uomo-internet (web mining).

La ricerca metodologica verrà condotta in un'ottica fortemente multidisciplinare, caratteristica di filoni di ricerca recenti quali il data mining ed il web mining. I gruppi di ricerca coinvolti nel progetto sono noti, a livello nazionale ed internazionale, per competenze specialistiche verticali su aspetti di tipo informatico, matematico, psicologico o statistico, ovvero dal punto di vista applicativo. Il gruppo di ricerca intende mettere in rete tali competenze e valorizzarne le sinergie in modo orizzontale.

La motivazione principale del progetto è la costruzione di un gruppo di ricerca completo, in grado di affrontare le diverse problematiche inerenti la ricerca e l'applicazione di metodi di data mining. A livello internazionale, il modello di riferimento è quello della ACM (Association for computing machinery), che ha un gruppo di ricerca interdisciplinare sul tema del Knowledge discovery in databases, spesso sinomimo di data mining e, a livello europeo, la rete KDnet. Ad oggi non risulta esser presente, nel nostro paese, un tale tipo di rete di ricerca.

Da un punto di vista più operativo, ci si occuperà di sei aspetti di ricerca, fra loro strettamente interconnessi, che verranno ora brevemente descritti.

Obiettivo 1: metodologie di web usage mining.
Il primo obiettivo riguarda la definizione di un repertorio di strumenti per l'analisi dei dati di e-business, attraverso lo sviluppo di metodologie per il web usage mining. Ci si propone in particolare lo studio dei percorsi di navigazione degli utenti di un sito, attraverso appropriati modelli statistici di tipo associativo.

Obiettivo 2: metodologie di web pattern discovery. Il secondo obiettivo si indirizza a problemi di recupero ed analisi dell' informazione, cruciali nel contesto dello sviluppo delle infrastrutture informative per l'e-business. Ci si propone di identificare nuove tecniche algoritmiche e combinatoriali alla base della scoperta automatica di pattern e delle loro "regole associative" in contesti e media disparati, e di specializzarne incarnazioni ad hoc nel settore dell' e-business.

Obiettivo 3: aspetti cognitivi nell'interazione uomo-macchina.
Il terzo obiettivo della ricerca è il raccordo fra la teoria delle decisioni, specie dal punto di vista cognitivo, ed il tema dell'interazione uomo-macchina. In particolare, il progetto si propone di indagare come vengono prese le scelte in ambienti online e da quali fattori esse sono influenzate.

Obiettivo 4: applicazione all'e-learning.
Il quarto obiettivo della ricerca prevede la progettazione di strumenti metodologici e la realizzazione di strumenti software a supporto per l'organizzazione dei contenuti dei siti di e-learning. In particolare, verrà fatto esplicito riferimento alla strutturazione di siti che erogano servizi e prodotti didattici.

Obiettivo 5: applicazione al customer relationship management.
Il quinto obiettivo della ricerca riguarda l'applicazione di tecniche di classificazione e di web usage mining nell'ambito del

marketing relazionale, e prevede lo sviluppo di modelli, metodi e strumenti per l'ottimizzazione delle azioni di marketing.

Obiettivo 6: applicazione all' e-government. Il sesto obiettivo della ricerca prevede l'applicazione delle metodologie di data mining a problemi di scelta delle decisioni in ambito pubblico.L'E-government è l'impiego di tecnologie di informazione e comunicazione nelle pubbliche amministrazioni, combinati con cambiamenti organizzativi e nuove abilità, al fine di migliorare la qualità dei servizi pubblici. Il lavoro di ricerca si propone di contribuire a tale obiettivo mediante l'impiego di modelli di data mining.

Testo inglese
The research project is focused on the study, the development and the implementation of the most appropriate data mining methodologies for e-business applications. The main emphasis of the research will be on the methodological and the applied aspects involved in the analysis of data arising from man-internet interactions (web mining).

The methodological research of the group will be carried out by a rather interdisciplinary team, as is typical of the research in very recent areas such as data mining and web mining. The applied research will mainly involve three e-business applications: the design of the architecture and of the contents of e-learning web sites; the application of web mining to customer relationship management problems; the management and the evaluation of public services, using information technology tools (e-government). The research groups that participate in the project are well known, at the national and international level, for vertical competences on some of the six research areas of the project; our aim is therefore to network such competences in an horizontal way, so to have cross-fertilisation of research competences and achievements.

The main motivation of the research project is to build a complete research network on the theme of web mining, at the Italian level. At the international level, our reference model is the ACM (Association for computing machinery) research group on Knowledge discovery in databases; at the European level the network of excellence Kdnet. A number of researchers in our group belong to such networks. To date we are not aware of similar research groups in Italy. From this viewpoint, our team is a blend of mature and competent researchers with young Phd and research assistants, still in a learning phase. We believe that this constitutes an ideal environment for the creation of a strong research network.

In operational terms, the research work will involve six research areas, strongly interconnected, that we now briefly describe.

Objective 1: web usage mining methodologies.
The first objective concerns the development of tools for e-business data analysis and, specifically, for web usage mining data. With such methodologies we aim to understand the navigation patterns of the visitors, in terms of appropriate association structures.

Objective 2: web pattern discovery methdologies. The second objective addresses issues of information retrieval and analysis that are perceived as crucial in the development of advanced infrastructures for the e-business . The project seeks to identify novel techniques supporting the automated discovery of patterns and their associations or ``rules'' in disparate contexts and media, and to fine tune their ad hoc incarnations in diverse fields.

Objective 3: cognitive aspects in man-internet interaction.
The third objective (M3) is the integration of decision theory with human-machine interaction research. Specifically, the aim of the project is to investigate how users make decisions into online environments and the factors that influence them.

Objective 4: application to e-learning.
The fourth objective of the research is aimed at developing the theory and at building software tools for the management of the contents of e-learning sites. In particular we shall be concerned with sites that erogate teaching services.

Obiettivo 5: application to customer relationship management.
The fifth objective concerns the development and the application of classification and web usage mining tools in the context of customer relationship management. It also concerns the development of models, methods and tools to improve marketing strategies.

Obiettivo 6: application to e-government. The sixth objective of the research is to employ data mining methods to aid public decision-making, in particular in the context of the evaluation of public services.

1.4 Durata del Programma di Ricerca

24 Mesi

1.5 Settori scientifico-disciplinari interessati dal Programma di Ricerca

SECS-S/01 - Statistica
ING-INF/05 - Sistemi di elaborazione delle informazioni
MAT/09 - Ricerca operativa
M-PSI/01 - Psicologia generale
SECS-S/06 - Metodi matematici dell'economia e delle scienze attuariali e finanziarie

1.6 Parole chiave

Testo italiano
DATA MINING ; WEB USAGE MINING ; WEB PATTERN DISCOVERY ; TEORIA DELLE DECISIONI ; E-LEARNING ; MARKETING RELAZIONALE ; E-GOVERNMENT ; MODELLI ASSOCIATIVI ; MODELLI DI CLASSIFICAZIONE

Testo inglese
DATA MINING ; WEB USAGE MINING ; WEB PATTERN DISCOVERY ; DECISION THEORY ; E-LEARNING ; CUSTOMER RELATIONSHIP MANAGEMENT ; E-GOVERNMENT ; ASSOCIATION MODELS ; CLASSIFICATION MODELS

1.7 Coordinatore Scientifico del Programma di Ricerca

GIUDICI ***PAOLO***

Professore Associato ***23/03/1965*** ***GDCPST65C23I829X***

SECS-S/01 - Statistica

Università degli Studi di PAVIA

Facoltà di ECONOMIA

Dipartimento di ECONOMIA POLITICA E METODI QUANTITATIVI

0382506224	***0382304226***	***giudici@unipv.it***
(Prefisso e telefono)	*(Numero fax)*	*(Email)*

1.8 Curriculum scientifico

Testo italiano
Paolo Giudici (Msc in Statistica, Minesota, 1989; Phd in Statistica, Trento, 1993) dal 1993 è ricercatore e dal 2000 Professore Associato di Statistica presso la Facoltà di Economia dell'Università di Pavia. È membro del comitato tecnico ordinatore del corso di laurea specialistico interfacoltà in "Management e Tecnologie dell'e-business" e responsabile di Facoltà per l'organizzazione di corsi post-laurea in modalità E-learning. È autore di circa 65 pubblicazioni scientifiche, delle quali due libri di ricerca, 30 articoli scientifici apparsi su riviste internazionali ISI e 33 articoli in atti di convegno e volumi sottoposti a referaggio. Gli argomenti di ricerca trattati hanno riguardato, in particolare, lo sviluppo di modelli statistici per il data mining, la statistica bayesiana ed i metodi computazionali di Markov Chain Monte Carlo. Ha costituito nel 2001 il laboratorio di data mining dell'Università di Pavia che svolge ricerca in collaborazione con enti di ricerca ed aziende, nazionali ed internazionali.
Il coordinatore ha partecipato attivamente a reti Europee di eccellenza, in particolare "Highly structured stochastic systems" (European Science Foundation, 1994-2000); "Computational and statistical methods for the analysis of spatial data" (EU Training and Mobility of Researchers, 1997-2001). Attualmente è responsabile dell'unità di ricerca di Pavia di "European Network for Promoting Business and Industrial Statistics" (PRO-E.N.B.I.S., EU Competitive and Sustainable Growth programme, 2003-2004).
E' stato inoltre invitato a presentare relazioni presso qualificate istuzioni di ricerca internazionali, fra le quali: University of Washington, Seattle; Isaac Newton Institute, Cambridge; The Fields Institute, Toronto; Trinity College, Dublin; Max-Planck Institute for Physics, Munich; Royal Statistical Society, London; SAMSI Institute, Duke University; University of Leuven; ETH Zentrum, Zurich.

Testo inglese
Paolo Giudici (Msc in statistics, Minnesota, 1989; Phd in statistics, Trento, 1993) is Associate Professor of Statistics at the Fcaulty of Economics of the University of Pavia. He is member of the board of a Master's degree on "E-business Managment and Technologies" and responsible for the E-learning activities of the Faculty of Economics of the University of Pavia.
He has authored about 65 scientific papers, among which 2 research books, 30 articles appeared in international journals (ISI) and 33 papers in refereed proceedings and volumes. His research themes can be classified into: statistical models for data mining, multivariate graphical models, bayesian statistics and Markov Chain Monte Carlo computational methods.
In 2001 he has founded the data mining laboratory of the University of Pavia that carries out research, applied and foundational, in collaboration with research institutions and companies.
The coordinator has actively participated in European excellence networks, such as: "Highly structured stochastic systems" (European Science Foundation, 1994-2000); "Computational and statistical methods for the analysis of spatial data" (EU Training and Mobility of Researchers, 1997-2001). Currently he is coordinator of the Pavia RU of the "European Network for Promoting Business and Industrial Statistics" (PRO-E.N.B.I.S., EU Competitive and Sustainable Growth programme, 2003-2004); he is an individual member of the "European Knowledge discovery network" (KDnet).
He has been invited to discuss his research work in a number of qualified research Institutions, among which: University of Washington, Seattle; Isaac Newton Institute, Cambridge; The Fields Institute, Toronto; Trinity College, Dublin; Max-Planck Institute for Physics, Munich; Royal Statistical Society, London; SAMSI Institute, Duke University; University of Leuven; ETH

Zentrum, Zurich.

1.9 Pubblicazioni scientifiche più significative del Coordinatore del Programma di Ricerca

1. *GIUDICI P. (2003).* ***Applied data mining: statistical methods for business and industry*** *pp. 1-364 ISBN: 0-470-84678-X LONDRA: Wiley (UNITED KINGDOM)*
2. *GIUDICI P.; BROOKS S.; GIUDICI P. (2003).* ***Efficient construction of reversible jump MCMC proposal distributions*** *JOURNAL OF THE ROYAL STATISTICAL SOCIETY. (vol. 65 pp. 3-55)*
3. *GIUDICI P.; CASTELO R. (2003).* ***Improving MCMC model search for data mining*** *MACHINE LEARNING. (vol. 50 pp. 127-158)*
4. *GIUDICI P.; CASTELO R. (2001).* ***Association models for web mining*** *DATA MINING AND KNOWLEDGE DISCOVERY. (vol. 5 pp. 183-196)*
5. *GIUDICI P. (2001).* ***Bayesian data mining, with application to credit scoring and benchmarking.*** *APPLIED STOCHASTIC MODELS FOR BUSINESS AND INDUSTRY. (vol. 17 pp. 69-81)*

1.10 Elenco delle Unità di Ricerca

nº	Responsabile Scientifico	Qualifica	Settore Disc.	Università	Dipartimento	Mesi Uomo
1.	*CARDACI MAURIZIO*	*Professore Ordinario*	*M-PSI/01*	*PALERMO*	*PSICOLOGIA*	*12*
2.	*GIUDICI PAOLO*	*Professore Associato*	*SECS-S/01*	*PAVIA*	*ECONOMIA POLITICA E METODI QUANTITATIVI*	*12*
3.	*GUERRA CONCETTINA*	*Professore Ordinario*	*ING-INF/05*	*PADOVA*	*INGEGNERIA DELL'INFORMAZIONE*	*12*
4.	*LA TORRE DAVIDE*	*Professore Associato*	*SECS-S/06*	*MILANO*	*ECONOMIA POLITICA E AZIENDALE*	*16*
5.	*SCHOIER GABRIELLA*	*Professore Associato*	*SECS-S/01*	*TRIESTE*	*SCIENZE ECONOMICHE E STATISTICHE*	*12*
6.	*VERCELLIS CARLO*	*Professore Ordinario*	*MAT/09*	*Politecnico di MILANO*	*INGEGNERIA GESTIONALE*	*16*

1.11 Mesi uomo complessivi dedicati al programma

		Numero	Mesi uomo 1° anno	Mesi uomo 2° anno	Totale mesi uomo
Personale universitario dell'Università sede dell'Unità di Ricerca		*25*	*124*	*124*	*248*
Personale universitario di altre Università		*6*	*33*	*26*	*59*
Titolari di assegni di ricerca		*0*			
Titolari di borse	*Dottorato*	*8*	*35*	*37*	*72*
	Post-dottorato	*0*			
	Scuola di Specializzazione	*0*			
Personale a contratto	*Assegnisti*	*2*	*17*	*17*	*34*
	Borsisti	*2*	*12*	*12*	*24*
	Dottorandi	*0*			
	Altre tipologie	*1*	*10*	*10*	*20*
Personale extrauniversitario		*1*	*6*	*6*	*12*
TOTALE		**45**	**237**	**232**	**469**

2.1 Obiettivo del Programma di Ricerca

Testo italiano
Gli obiettivi principali del progetto di ricerca possono essere suddivisi in sei aree, strettamente interconnesse ed aventi in comune la necessità di strumenti di data mining adeguati.

Il primo obiettivo di ricerca è lo sviluppo di metodi per l'analisi dei dati di visita ad un sito web (web usage mining). Verranno confrontati e sviluppati modelli associativi per l'individuazione dei percorsi di visita più probabili. Nella letteratura corrente tali metodologie fanno prevalentemente riferimento alle regole associative e sequenziali, implementate, ad esempio, nell'algoritmo a priori. Tali regole associative, sebbene facilmente implementabili in algoritmi che ne permettono la ricerca efficiente delle più rilevanti, hanno come limite intrinseco il fatto di essere calcolate in modo locale, ovvero, dal punto di vista statistico, marginale, senza considerare le interdipendenze fra tutte le variabili presenti nel database. Inoltre tali regole non considerano gli aspetti inferenziali. L'obiettivo della ricerca sarà il confronto (statistico e computazionale) delle regole associative, di tipo "locale", con modelli statistici di tipo "globale": multivariati e corredati da aspetti inferenziali, da sviluppare a partire da classi di modelli statistici noti in letteratura, quali i modelli grafici multivariati, le reti bayesiane e le catene di Markov. Un ulteriore obiettivo della ricerca sarà lo sviluppo di metodologie bayesiane per l'analisi dei dati di web usage mining, e l'implementazione dei corredati algoritmi simulativi di Markov Chain Monte Carlo per l'effettuazione delle inferenze di interesse (scelta del modello, stima dei parametri).

Il secondo obiettivo della ricerca riguarda il web pattern discovery. La individuazione di pattern è basata su proprietà strettamente intercorrelate della statistica, del confronto di forme e della combinatorica su parole. Queste proprietà consentono di limitare drasticamente ed a priori l'insieme dei pattern candidati sovra-rappresentati o sotto-rappresentati per tutte le lunghezze in una data sequenza, rendendo quindi possibile sia la individuazione che la visualizzazione di tali pattern in modo pratico e veloce. L' approccio proposto consiste della annotazione di un indice quale un albero dei suffissi o automa delle sottosequenze. Sotto una varietà di assunzioni probabilistiche, si è trovato che tali annotazioni possono essere effettuate in modo efficiente in termini di spazio e tempo per la media, la varianza, ed alcune altre misure di significatività comunemente adottate, anche senza imporre restrizioni sulla lunghezza dei pattern esaminati. Il lavoro teorico che deve essere fatto consiste nella estensione di questi costrutti a modelli probabilistici meno semplicistici di quelli già trattati. In pratica, proponiamo di integrare questi fatti in uno strumento software, adattandolo a soddisfare in modo unificato un ampio reportorio di richieste.
Saranno anche investigate tecniche di indicizzazione multi-dimensionale per l'organizzazione, l'accesso e la integrazione di dati di diversa natura. La ricerca proposta si pone come obiettivo anche quello di mettere a punto schemi di indici multidimensionali efficienti sulla base di proprieta' statistiche dei dati disponibili.

Nel contesto del terzo obiettivo di ricerca, sottolineiamo che la ricerca sulla presa di decisione è stata finora condotta prevalentemente in ambienti offline. Data l'attuale espansione di Internet e degli spazi di compravendita online si impone la necessità di estendere lo studio relativo alle scelte d'acquisto al mondo virtuale. In quest'ultimo, l'utente si trova a dover confrontare tra loro una miriade di alternative descritte lungo un'infinità di attributi, disponendo tuttavia di capacità cognitive limitate. A fronte di tale difficoltà, diviene fondamentale comprendere come l'informazione debba essere presentata nei siti e-commerce per facilitare i processi decisionali. Si esplorerà il ruolo giocato dalla specifica composizione del set di scelta sulle preferenze degli utenti, focalizzandosi sulle relazioni di dominanza esistenti fra le varie alternative proposte. Di tale effetto, noto come 'attraction effect', ci si propone di esplorarne le ricadute negli ambienti online, al fine di ottenere utili suggerimenti per la messa a punto di nuovi siti di commercio elettronico. Proprio su tale versante applicativo si sta orientando la ricerca più avanzata sul web mining. Attualmente si assiste, infatti, ad un continuo diffondersi di molteplici siti Web, i cosiddetti Decision-Facilitating-Websites, o Siti d'Aiuto alle Decisioni (S.A.D.), sorti inizialmente negli Stati Uniti, ma attualmente presenti anche in Europa, al fine di consentire ai navigatori di reperire velocemente le informazioni desiderate, integrarle e prendere in ultimo decisioni ottimali, con il minimo sforzo possibile.

Il quarto obiettivo condivide le motivazioni di fondo del terzo obiettivo e le contestualizza nell'ambito dell'E-learning. Dal punto di vista dell'autore di corsi di E-learning, concepiti come opportuna composizione di learning object già esistenti e reperibili da diverse fonti, diventa molto importante poter e saper selezionare gli oggetti informativi che meglio si adattano al particolare contesto. Spesso questo processo decisionale è lasciato alla pura preferenza dell'autore, senza che si possa tener conto dei fattori oggettivi (quasi sempre a prima vista non evidenti) che differenziano i diversi learning object tra loro. Un adeguato sistema di supporto basato su data mining diventa allora importante in tutti quei casi in cui la garanzia di successo del processo formativo a distanza rappresenti un requisito irrinunciabile.

Il quinto obiettivo di ricerca è inerente al marketing relazionale e si rivolgerà ai seguenti obiettivi specifici: sviluppo di metodi rivolti all'analisi delle e-mail e delle clickstreams, e all'integrazione di queste informazioni con le più tradizionali informazioni demografiche e transazionali, per sviluppare profilazioni e segmentazioni dei clienti; sviluppo di modelli di ottimizzazione multi-periodo per massimizzare la redemption e il ritorno sull'investimento delle campagne di marketing. L'attività di ricerca rivolta alle applicazioni di marketing relazionale richiederà soprattutto di sviluppare nuovi modelli matematici di classificazione, con particolare riferimento alle estensioni discrete delle support vector machines. Gli specifici ambiti di indagine considerati riguarderanno: metodi DSVM (discrete support vector machines) e SVM basati su kernels nonlineari; metodi DSVM e SVM per la classificazione multi-categorica; metodi DSVM e SVM per la classificazione di testi; sviluppo di criteri di misura e benchmarking per il confronto tra metodi di classificazione; metodi di selezione degli attributi in presenza di informazioni strutturate e semi-strutturate; metodi di classificazione di tipo PET (probability estimation trees) per l'assegnazione di punteggi individuali alle istanze di un dataset.

Infine, il sesto obiettivo della ricerca prevede l'applicazione delle metodologie di data mining a problemi di scelta delle decisioni in ambito pubblico. In tale ambito, l'E-government è definito come l'uso di tecnologie di informazione e comunicazione nelle pubbliche amministrazioni, combinati con cambiamenti organizzativi e nuove abilità, al fine di migliorare i servizi pubblici, i processi democratici e rafforzare il supporto alle politiche pubbliche. I processi di cambiamento nell'organizzazione e nella cultura hanno bisogno di tempo: possono essere necessari diversi anni prima che un investimento fornisca benefici completi. L'obiettivo principale del gruppo di ricerca è quello di contribuire allo sviluppo dell'e-government mediante lo sviluppo e l'applicazione di modelli matematici e statistici di data mining.

Testo inglese
The aims of the proposed research can be grouped into six main areas, with many common aspects, such as the common need of developing adequate data mining tools.

The first research objective concerns the development of data mining methods for web usage mining. We intend to compare and develop association models to individuate the most likely navigation paths. In the current data mining literature, such methods mainly refer to association and sequence rules, and variants thereof. Such rules, although easy to interpret and implement (for instance, in the a priori algorithm) are calculated locally, that is, marginally, without taking into account multivariate dependences between the variables in the database. Besides, these rules do not consider inferential aspects. The aim of the research will be the comparison (in terms of both computational and statistical efficiency) between "local" association rules and "global" models, multivariate and endowed with a probability model, allowing an inferential analysis. The latter will be developed starting from statistical models known in the literature, such as graphical models, bayesian networks and Markov chains. A further objective of the research will be the development of Bayesian models for web usage mining, and of Markov Chain Monte Carlo algorithms (MCMC) necessary to draw inferences, such as model scores and model estimates.

The second objective of the research addresses issues of information retrieval and analysis that are perceived as crucial in the development of advanced infrastructures for the e-business . The project seeks to identify novel techniques supporting the automated discovery of patterns and their associations or ``rules'' in disparate contexts and media, and to fine tune their ad hoc incarnations in diverse fields. Pattern discovery is based on subtly interwoven properties of statistics, pattern matching and combinatorics on words. These properties enable one to limit drastically and a priori the set of over- or under-represented candidate patterns of all lengths in a given sequence, thereby rendering it more feasible both to detect and visualize such patterns in a fast and practically useful way. The approach proposed is that of annotating a suffix tree or a directed acyclic word graph of a string. It turns out that, under a variety of basic probabilistic assumptions, such annotations can be carried out in a time- and space-efficient fashion for the mean, variance and some of the adopted measures of significance, even without setting limits on the length of the patterns considered. More importantly, the candidate over- or underrepresented words that need to be computed are linear rather than quadratic in the length of the textstring. The theoretical work to be done consists of extending these constructs to handle less simplistic probabilistic models. In practice, we plan to integrate these facts in an extension of a software tool adapting it to fulfill in a unified fashion a wide repertoire of needs.

The third objective of the research is the integration of decision theory with human-machine interaction research. Specifically, the aim of the project is to investigate how users make decisions into online environments and the factors that influence them. Most of the previous studies on decision-making concern off-line choices. Considering the growth of the Internet and the E-commerce web-sites, the study on consumer choices should be extended to web-based shopping environments. In these virtual environments, users are forced to compare a multitude of alternatives defined by an infinite number of attributes, having limited cognitive abilities. Because of this difficulty, is important to understand the way in which the information should be presented in the E-commerce web-sites in order to facilitate the decisional processes. The role on the individual preferences played by the specific compositions of the choice set will be explored, focusing on dominance relations existing among the different alternatives proposed.
One of these effects, named 'attraction effect', will be studied in online environments, in order to suggest useful information to create new e-commerce facilitating web-sites. Towards this end is oriented the most advanced research on decision making and data mining. Today, indeed, one sees the rapid rise of certain Websites, known as Decision-Facilitating-Websites, with the aim to allow the users to quickly find the desired information, to integrate that information and make optimal decisions with the least possible effort. First seen in the U.S., these Websites are now becoming increasingly popular in Europe.

The fourth objective of the research shares the motivational background of the third objective. The specific research aims concern the device of new tools (dedicated Web sites) capable of facilitating the representation of decisional situations which are to be faced in the creation of "didactic paths" based on E-learning. Such aims will be pursued through the development of visual techniques presenting the user (the one who has to take decisions) with the information he or she needs (in a simple and effective way). For the author of E-learning courses, considered as proper compositions of already-existing "learning objects" (which can be obtained from different sources), it is extremely important to be able to select those information objects which are suitable for the particular context. Often this fundamental decisional process is left to the author, without considering objective factors (almost always "hidden") which differentiate the various learning objects. A proper support system based on data mining thus becomes important in all those cases in which the success of the E-learning process must be guaranteed.

The fifth objective of the research concerns the application of classification techniques and web usage mining to relational marketing activities, and consists of developing models, methods and tools supporting campaign management optimization. In particular, the research will focus upon the following specific topics: the analysis of clickstreams and e-mails and their subsequent integration with the more traditional demographic and transactional information in order to profile and segment the customer base; the development of new multi-period optimization models to maximize the overall return on the marketing investment. The research activity dealing with relational marketing applications will be mainly developed by developing new mathematical models for classification, mostly derived from discrete support vector machines extensions. The research will focus upon the following specific topics: DSVM (discrete support vector machines) and SVM methods based on nonlinear kernels; DSVM and SVM methods for multicategory classification; DSVM and SVM methods for text recognition; definition of measurement indexes to compare the performances among alternative classification approaches; feature selection and feature extraction methods for structured and semi-structured information; PET (probability estimation trees) to assign a score to each instance of a dataset.

The sixth objective of the research is concerned with the study of data mining techniques for e-government applications. E-Government is defined as the use of information and communication technology in public administrations combined with organizational change and new skills in order to improve public services and democratic processes and strengthen support to public policies. The main goal of the research group is to to develop mathematical and statistical models of data mining to improve e-government applications. In particular we shall develop innovative models to measure quality of public services, as possible alternatives to simpler methods borrowed from marketing applications, such as customer satisfaction and customer loyalty (life time value) models.

2.2 Base di partenza scientifica nazionale o internazionale

Testo italiano

Il primo filone di ricerca si soffermerà in modo particolare sullo sviluppo di metodi per l'analisi dei dati di web usage mining. In tale ambito si pone spesso il problema dello sviluppo di adeguate metodologie per l'individuazione della struttura associativa presente fra le (molteplici) variabili a disposizione.
La letteratura inerente l'argomento è prevalentemente concentrata sugli aspetti informatici connesse alle cosiddette regole associative, proposte da Agrawal et al. (1995). Si vedano a tal proposito le monografie di Baldi, Frasconi e Smyth (2003), Chakrabarti (2003), Han e Kamber (2001) e Hand, Mannilla e Smyth (2001).
Alcuni ricercatori del gruppo hanno recentemente sviluppato, in collaborazione con altri ricercatori, nazionali ed internazionali, delle metodologie originali di scelta della struttura associativa, basate su modelli statistici di analisi multivariata (modelli grafici, sistemi esperti probabilistici) oltre che sull'impostazione bayesiana e sull'utilizzo di metodi computazionali di Monte Carlo basati sulle catene di Markov. I risultati più rilevanti in tale ambito sono contenuti in Giudici, 2001, 2003; Giudici e Green, 1999; Giudici e Castelo, 2003; Brooks, Giudici e Roberts, 2003. Quanto proposto ha trovato le sue più importanti applicazioni nell'ambito del risk management (Giudici, 2001; Giudici e Cornalba, 2004), del web usage mining (Giudici e Schoier, 2002; Castelo e Giudici, 2001) e della market basket analysis (Giudici e Passerone, 2001).
Gli sviluppi precedenti sono collegati ad un'intensa attività di discussione degli stessi, in particolare nell'ambito di tre convegni internazionali organizzati localmente a Pavia. Si vedano, a tal proposito, i volumi a cura di Giudici e Polasek, 2001; Giudici, Heckerman e Whittaker, 2001; Giudici, 2002, che contengono una selezione dei lavori pubblicati a tali convegni.

Con riferimento al secondo obiettivo di ricerca, si verifica che con l'emergere di infrastrutture dell'informazione, vengono ammassati, distribuiti e condivisi a ritmi sempre maggiori volumi di informazioni di dimensioni senza precedenti. L'accesso e la manipolazione efficaci delle informazioni dipendono pertanto in modo cruciale dall'efficienza con cui le stesse verranno strutturate, compresse, trasmesse, archiviate e recuperate. In questo contesto, l'identificazione di regolarita' e anomalie gioca un ruolo crescente e cruciale nell'ambito generale dell'organizzazione e recupero dell'informazione.
A fronte di dati che vengono ammassati incessantemente, il problema prevalente e' divenuto quello di limitare e filtrare cio' che una interrogazione debba restituire, e pertanto come generare delle caratterizzazioni succinte ed evidenziare le caratteristiche principali dei dati a disposizione. La costruzione di ``metadati'' e' un compito intermedio essenziale al raggiungimento di tali obiettivi. Ma, soprattutto, il potere, in modo automatico, individuare o generare pattern e associazioni, diventera' gradualmente l'unico modo per accedere a dati e informazioni troppo imponenti per essere gestibili altrimenti. Recentemente, si e' ricorso ai termini Scoperta di Pattern e Scoperta di Regole, nel tentativo di incapsulare un repertorio di problemi sintattici e strumenti adatti all'identificazione di regolarita' quali ripetizioni, cadenze, motivi, e occorrenze congiunte o connesse degli stessi in alcuni oggetti discreti elementari. Predecessori naturali di tali problemi e metodologie si sono presentati gia' dagli anni '70 in contesti quali, ad es., il disegno di compilatori. Da allora, il raggio di applicazione per strumenti e metodiche di ``pattern matching'' si e' allargato all'elaborazione di testi, immagini e segnali, analisi e riconoscimento del parlato, compressione di dati, biologia computazionale, chimica computazionale, visione computerizzata, ecc. (Apostolico (2003), Apostolico, Bock e Lonardi (2003). A Apostolico e Atallah (2002) A. Apostolico and Z. Galil (1997), M. Crochemore and W. Rytter (1996). La scoperta di pattern e regole puo' essere vista come un insieme di problemi nuovi di pattern matching portati alla luce dal recente emergere esplosivo di applicazioni collegate ai sistemi multimediali, biologia molecolare computazionale, sistemi di grandi basi di dati, fornitura di informazioni su scala mondiale, e nuove tecnologie hardware e software. Con l'accumularsi di siti web e banche dati, si rendono necessari metodi sempre piu' veloci e sofisticati, in particolare relativamente a tutti gli stadi della ricerca, matching, confronto e analisi di strutture discrete elementari quali stringhe, alberi, vettori, espressioni regolari, alcune classi speciali di grafi, e loro combinazioni. Questi problemi costituiscono un territorio vasto ed inesplorato e presentano sia opportunita' che difficolta' considerevoli.

Nel contesto del terzo obiettivo di ricerca, si evidenzia come negli ultimi anni si sia assistito ad una crescente attenzione verso l'integrazione delle tecniche di data mining e della teoria della decisione. Grazie, infatti, alla diffusione delle nuove tecnologie digitali e al rapido incremento dell'uso di Internet è divenuto ormai possibile accedere a una miriade d'informazioni circa le opzioni di scelta e le loro caratteristiche. Tutto ciò, se da un lato rappresenta un considerevole vantaggio per l'utente, consentendo di compiere 'scelte informate', basate sulla valutazione di molteplici alternative, dall'altro lato, tuttavia, rappresenta un complicato problema, nella misura in cui l'individuo si trova a dover gestire una quantità sconfinata di informazioni, avendo però capacità cognitive limitate.
La letteratura classica sul decision making si è basata per lungo tempo sull'assunto secondo cui scegliere fra molte opzioni piuttosto che fra poche sia preferibile poiché motivazionalmente più stimolante (p.e. Deci, 1975; Deci e Ryan, 1985; Taylor, 1989). Tali studi hanno tuttavia sollevato non poche controargomentazioni in merito soprattutto all'artificiosità dei compiti e delle situazioni sperimentali utilizzate. In particolare, si deve a Iyengar e Lepper (2000) il merito di aver sottolineato la mancanza di validità ecologica degli esperimenti classici sul decision making. In questi, infatti, i soggetti venivano invitati a scegliere fra un certo numero di opzioni che variava all'interno di una gamma ristretta, in particolare fra due (poche opzioni) a sei (considerate 'molte'opzioni). E' ovvio che in simile situazione decidere fra molte alternative risulta nettamente preferibile che scegliere fra sole due. Appare tuttavia immediatamente evidente che tali set decisionali non siano rappresentativi delle complesse situazioni di scelta che caratterizzano la realtà quotidiana in cui il numero di alternative è considerevolmente più elevato (si pensi a titolo esemplificativo alle scelte operate su Internet).

A partire da tale constatazione Iyengar e Lepper (2000) conducono una ricerca all'interno di un comune supermercato, in cui compito dei soggetti è scegliere fra un numero elevato di alternative, questa volta 23, e un numero ristretto, soltanto 6. Contrariamente all'assunto classico della Psicologia della decisione i risultati dimostrano che scegliere fra un numero esteso di alternative, sebbene inizialmente possa sembrare maggiormente desiderabile, durante l'intero processo decisionale determina un forte effetto demotivante. All'interno di tale quadro di riferimento teorico assumiamo che il numero delle opzioni di per sé non sia sufficiente per determinare il grado di difficoltà della scelta. Molto dipende, infatti, dalla specifica relazione esistente fra le varie alternative decisionali. Come dimostrano gli studi compiuti negli ultimi venti anni sui fattori situazionali del decision making, la scelta è sostanzialmente influenzata da specifici effetti del contesto attribuibili alla particolare composizione del set decisionale. In particolare, le ricerche sull'effetto attraction (Huber, Payne e Puto, 1982) evidenziano come le relazioni di dominanza esistenti fra le opzioni influenzino sistematicamente la scelta. Alla luce di tale dato, il presente progetto si pone l'obiettivo di verificare se le relazioni di dominanza influenzino la scelta fra poche vs molte opzioni. Ci si propone di condurre il presente esperimento in un ambiente virtuale, come quello tipico del commercio elettronico.

Quanto appena affermato può essere contestualizzato all'ambito dell'e-learning, a cui fa riferimento il quarto progetto di ricerca (si vedano ad esempio Spence, 2001 e Furnas, 1997). Grazie alle tecnologie digitali ed al rapido incremento dell'uso di Internet è divenuto ormai possibile accedere ad una grande quantità di informazioni circa i prodotti disponibili (nella fattispecie, "servizi" didattici, quali ad esempio piattaforme, corsi, learning object, ecc.) e le loro caratteristiche (si veda ad esempio il rapporto LearnFrame, 2000). Per l'autore di corsi di E-learning, concepiti come opportuna composizione di "contenuti elementari" (i learning object) già esistenti e reperibili da diverse fonti, diventa importantissimo poter e saper selezionare gli oggetti informativi che meglio si adattano al particolare contesto. Spesso questo fondamentale processo decisionale è lasciato alla pura preferenza dell'autore, senza che si possa tener conto dei fattori oggettivi (quasi sempre a prima vista non evidenti) che differenziano i diversi learning object tra loro (si veda Cantoni, Cellario e Porta, 2003).
Nel campo dell'E-learning, un difficile problema di scelta si presenta al "creatore di percorsi didattici" (l'autore dei corsi), nel momento in cui si trova a dover scegliere tra più alternative nell'uso di "moduli" già pronti e riutilizzabili, ossia i cosiddetti Learning Object (LO). Anche se le possibili definizioni sono molte, un LO si può vedere come un "blocco autosufficiente" di contenuto informativo che soddisfi un ben preciso "obiettivo di apprendimento". Tra i vantaggi derivanti dalla strutturazione dei contenuti in LO, uno dei principali risiede sicuramente nella riduzione dei costi di creazione dei corsi, garantita dal riuso di componenti già esistenti e dal fatto che diverse parti possono spesso essere mantenute e aggiornate separatamente. L'adozione ormai quasi universale di standard (quali SCORM) per la metadatazione dei LO favorisce questo processo di "riuso" di materiale esistente e garantisce la compatibilità anche tra piattaforme di diversi produttori. Non va poi dimenticato che sempre più si sente la necessità di adattare il percorso formativo alle esigenze del singolo utente; un buon sistema di gestione del processo di apprendimento (Learning Management System, LMS) dovrebbe allora essere in grado di identificare il "livello" dell'utente, creando quindi un percorso personalizzato. Sono ormai disponibili numerosi database (Digital Repository, DR) che raccolgono LO riutilizzabili a seconda delle esigenze per costruire propri corsi. Anche se allo stato attuale non tutti i DR lo rispettano, esiste anche uno standard (Digital Repository Interoperability, DRI) che rende ad esempio possibile eseguire ricerche multiple di LO su più DR. Ciò vuol dire che il creatore di percorsi didattici sempre più in futuro si troverà a dover scegliere tra moltissime possibilità, che possono differire tra loro anche per molti fattori. E' quindi importante disporre di sistemi di data mining che facilitino il processo decisionale, attraverso opportune interfacce che permettano di analizzare e confrontare le caratteristiche salienti dei LO. Questo è l'obiettivo della ricerca proposta.

Nel contesto del quinto obiettivo di ricerca, sottolineiamo che le tecniche di data mining (Berry e Linoff, 1997) si sono recentemente proposte come uno strumento fondamentale a supporto di attività di marketing, quali il marketing one-to-one (Peppers e Rogers, 1993) e il customer relationship management (CRM) (Brown, 2000). Il tradizionale approccio di tipo mass-market basato sull'interazione simultanea con un numero elevato di clienti, tramite canali di comunicazione quali TV o giornali, è stato progressivamente sostituito da azioni di marketing mirate, con l'obiettivo di rispondere puntualmente ai bisogni dei clienti. L'avvento dell' e-commerce e la possibilità di registrare vaste moli di dati transazionali hanno drasticamente accelerato l'applicazione delle metodologie di data mining al marketing relazionale. Accanto ai dati demografici e transazionali, anche le informazioni relative alle interazioni on-line stanno acquisendo un'importanza crescente. In questo contesto, le informazioni più rilevanti derivano da attività di clickstream analysis, nonché dalle interazioni che occorrono tramite sistemi wireless, TV via cavo, e-mails. Nel marketing relazionale si possono identificare quattro principali ambiti di analisi: la conquista di nuovi clienti; la fidelizzazione dei clienti; le attività di cross-selling e up-selling. Tutti sono riconducibili al paradigma logico della classificazione. Nell'ambito della classificazione, una delle tecniche di maggior successo è costituita dalle support vector machines (SVM), proposte da Vapnik (1995, 1998) e basate sul principio di minimizzazione del rischio strutturale: tale principio formalizza l'esigenza di ridurre simultaneamente l'errore empirico e l'errore di generalizzazione al fine di conseguire predizioni più accurate (Cristianini e Shawe-Taylor, 2000)(Schölkopf e Smola, 2002). Dalla formulazione classica delle SVM è stato di recente sviluppato in diverse varianti il concetto di discrete support vector machines (DSVM) per problemi di classificazione binaria (Orsenigo e Vercellis, 2003a, 2004b, 2004c; La Torre e Vercellis, 2003). Il paradigma delle DSVM è stato recentemente esteso a problemi di classificazione multicategorica (Orsenigo and Vercellis, 2003b). L'ottimizzazione delle campagne si propone di determinare la combinazione più vantaggiosa di offerte, segmenti di clienti e canali di comunicazione, nel rispetto dei vincoli riguardanti l'attività promozionale (Vercellis, 2002).

Infine, con riferimento al sesto progetto di ricerca, sottolineiamo che l'erogazione e la valutazione dei servizi pubblici (in ambito sanitario, scolastico, energetico, ambientale, trasporti) richiedono più che mai oggi analisi fondate su basi scientifiche. In generale le scelte vengono effettuate senza preliminari studi scientifici che garantiscano la massimizzazione dei benefici e la minimizzazione dei costi. Nella maggior parte dei casi finisce per prevalere un obiettivo sugli altri, spesso solo quello legato al profitto, quando invece la scelta di come, quanto e dove erogare un servizio dovrebbe essere il miglior compromesso tra una serie di obiettivi antitetici. L'E-government è definito come l'uso di tecnologie di informazione e comunicazione nelle pubbliche amministrazioni combinati con cambiamenti organizzativi e nuove abilità al fine di migliorare i servizi pubblici, i processi democratici e rafforzare il supporto alle politiche pubbliche (Beckford 2001, Borroughs et al 2000, Bowerman 2002, DETR 1998, Klazinga 2000, McIver 1998, Packwood et al 1998, Pollitt 1996). L'obiettivo principale della ricerca è quello di usare sofisticati modelli matematici e statistici di data mining per migliorare la valutazione e la distribuzione dei servizi pubblici. La qualità di un servizio è un comune argomento di ricerca nel management e nel marketing dalla fine del 1970 (Williams, B. 1994, Cabinet Office 1999, Campbell, S. M.

et al 2000, Freeman, T. 2002). La natura della qualità di un servizio è differente dalle qualità di beni fisici ed è fortemente relazionata con i concetti tradizionali di soddisfazione del cliente e lealtà del cliente. Dal 1980 i ricercatori hanno studiato diversi modelli per l'analisi della qualità di un servizio. Ci proponiamo di confrontare questi modelli con nuove metodologie di valutazione dei servizi pubblici, basate su modelli di data mining.

Testo inglese

The first research objective will concern the development of data mining methodologies for web usage mining. In this context a very important open problem is the determination of the association structure present among the many variables in the database (logfile). The available literature on the topic is mainly focused on association and sequence rules, proposed by Agrawal et al. (1995), with the corresponding search algorithm, the so-called a priori algorithm. For a detailed review see the monographs by Baldi, Frasconi and Smyth (2003), Chakrabarti (2003), Han and Kamber (2001) and Hand, Mannilla and Smyth (2001).
The recent activity of some researchers in the group has recently turned to the development of statistical models to detect associations in large databases. Such models have been based on multivariate statistical models (graphical models, expert systems and Markov chains) and/or on Bayesian modelling, aided by Markov Chain Monte Carlo methods for the approximation of the required inferences.
The most relevant results in this context have been obtained in
Giudici, 2001, 2003; Giudici and Green, 1999; Giudici and Castelo, 2003; Brooks, Giudici and Roberts, 2003. The suggested methods have been applied to risk management (Giudici, 2001 and Giudici and Cornalba, 2004), Web usage mining(Castelo and Giudici, 2001, Giudici and Schoier, 2002) and market basket analysis (Castelo and Giudici, 2001).
The above results have been developed in connection with an intense dissemination activity, particularly in three organised conferences on data mining. The volumes edited by Giudici and Polasek, 2000; by Giudici, Heckerman and Whittaker, 2001; and Giudici, 2002 contain a selection of the papers presented at such conferences. The aim of the research is to consolidate such a research netowrk and improve research in association models for web mining, as detailed later in the description of the planned research activities.

As for the second objective of the research, it turns out that in the emerging information infrastructures, unprecedented volumes of information are amassed, disseminated and shared at an increasing pace. Effective access to, and manipulation of information depends thus crucially on the efficiency with which information itself is structured, compressed, transmitted, stored and retrieved. In this context, the detection of regularities and anomalies plays a crucial role for the general organization and retrieval of information.
With data being unendingly amassed, the prevailing problem has become one on how to limit and filter what a query shall return, hence to generate succinct characterizations and enhance prominent features for the available data. Metadata buildup is an essential intermediate task towards these objectives. But above all, the ability to automatically detect or generate patterns and associations will gradually become the only means of access to data and information too huge to be palatable. The terms Pattern Discovery and Rule Discovery begun to be used recently in an attempt to encapsulate a repertoire of syntactic problems and tools akin to the identification of regularities such as repetitions, cadences, motifs, and joint or connected occurrences thereof in some elementary discrete objects. The natural predecessors of such problems and techniques spun as early as in the 70's in contexts such as, e.g., compiler design. Since then, the range of applications of "pattern matching" tools and methods has spawned to text, image and signal processing, speech analysis and recognition, data compression, computational biology, computational chemistry, computer vision, etc. . (Apostolico (2003), Apostolico, Bock e Lonardi (2003). A Apostolico e Atallah (2002) A. Apostolico and Z. Galil (1997), M. Crochemore and W. Rytter (1996). Discovery may be regarded as a set of novel pattern matching problems brought about by the recent explosive emergence of applications related to multimedia systems, computational molecular biology, very large database systems, worldwide information servers, and new software and hardware technology. As WEB sites and data banks accumulate, increasingly fast and sophisticated methods are sought, in particular, in connection with all stages of searching, matching, comparing and analyzing elementary discrete structures such as strings, trees, arrays, regular expressions, some special classes of graphs, and compounds thereof. These problems constitute a largely unexplored territory and pose both considerable opportunities and challenges.

Referring to the third objective of the research, the integration of data mining techniques with decision theory has received increasing attention in recent years. In fact, the expansion of new digital techologies and the rapid increase in the use of the Internet has made it possible to access a multitude of information regarding choice options and their features. On one hand, this represents a considerable advantage, allowing one to make'informed choices' based on the evaluation of various alternatives. On the other hand, it represents a complex problem, because the user must manage an enormous amount of information while still bound by limited cognitive abilities. Because of this limitation, is important to understand how the information has to be showed on the Web to facilitate the difficult choice process. The classical literature on decision making is based on the assumption that having more rather than fewer choices is more desiderable and intrinsically motivating (e.g. Deci, 1975; Deci & Ryan, 1985; Taylor, 1989). The above studies have been criticized for the artificial tasks and experimental procedures. Specifically, Iyengar and Lepper (2000) underlined the lack of ecological validity of the classical experiments on decision making. Indeed, in these studies participants were invited to choose among a small number of alternatives, tipically between two (few options) and six (considered as 'many' options). Obviously, in this case, deciding among many alternatives is preferable to deciding between only two options. However, it is immediately clear that the above decisional set is not at all representative of the complex real-world situations in which the number of alternatives is considerably greater (for example, the choices available on the Internet).
Starting from these remarks, Iyengar and Lepper (2000) conducted research in a grocery store in which participants made a decision among 6 (limited-choice condition) or 24(extensive-choice condition) options. In contrast with the traditional models, the findings from this study showed that while having more choices at first appears more desiderable, subsequently it yields a strong demotivating effect.
In this theoretical framework we assume that the number of options alone is not enough to determine the difficulty of the choice. It depends, indeed, on the specific relation existing among the various alternatives. As demonstrated by studies over the last 20 years on situational factors of decision making, choice is substantially influenced by specific context effect as the composition of the decision set. In particular, the investigation of the attraction effect (Huber, Payne & Puto, 1982) showed that the relations of dominance among the options systematically influenced the choice. In light of this result, the aim of this project is to verify if the dominance relations influence the choice between few vs. many options. The experiment will be realized in a virtual environment, as that typical of E-commerce.

The previous consideration also apply to the fourth objective of the research, concerning e-learning, where the main aim of our research is to provide useful hints to devise new tools (dedicated Web sites) capable of facilitating the representation of decisional situations which are to be faced in the creation of "didactic paths" based on E-learning (Cantoni, Cellario and Porta, 2003). The objective will be pursued through the development of visual techniques presenting the user (the one who has to take decisions) with the information he or she needs (in a simple and effective way). See, for instance, Furnas, 1997 and Spencer, 2001. Thanks to digital technologies and to the rapid growth of Internet usage, it is now possible to access a great amount of information on the available products (in our case, "didactic services" such as E-learning platforms, courses, learning objects, etc.) and their characteristics (see, for instance, Learnframe, 2000). For the author of E-learning courses, considered as proper compositions of already-existing "learning objects" (which can be obtained from different sources), it is extremely important to be able to select those information objects which are suitable for the particular context. Often this fundamental decisional process is left to the author, without considering objective factors (almost always "hidden") which differentiate the various learning objects. A proper support system based on data mining becomes then important in all those cases in which the success of the E-learning process must be guaranteed. A difficult decision making problem must be faced by the "content creator", when he or she has to choose among several alternatives in selecting ready-to-use and reusable modules, i.e. the so-called Learning Objects (LOs). Even if there are several possible definitions, a LO can be seen as a self-contained block of learning which fulfils a single, stated "learning objective". Among the advantages deriving from organizing the content through LOs, one of the most important is the reduction of creation costs, made possible by the reuse of already-existing components. The (almost) universal adoption of standards (such as SCORM) to describe LOs through metadata promotes this "recycle process" of existing material and guarantees the compatibility with platforms of different vendors. Also, it is more and more important to adapt "learning paths" to the needs of the specific user; a well-designed system for the management of the learning process (Learning Management System) should be able to identify the user's "level", and create a personalized path. Several databases are now available (Digital Repositories, DRs) which contain LOs that can be reused according to specific requirements to build courses. Even though, currently, not all DRs respect it, a standard there exists (Digital Repository Interoperability, DRI) which makes it possible to perform multiple searches of LOs over several DRs. This means that the "content creator of the future" more and more will have to choose among many possibilities, which may differ in many aspects. It is therefore important to use systems that facilitate the decision process, through proper interfaces allowing the user to analyze and compare the main characteristics of LOs.

Referring to the fifth objective of the research (A2), we observe that data mining techniques (Berry and Linoff, 1997) are emerging as a key tool for implementing advanced marketing approaches, such as one-to-one marketing (Peppers and Rogers, 1993) or customer relationship management (CRM) (Brown, 2000). The traditional mass market approach based on interactions with a large number of customers simultaneously, using broadcast channels such as TV or magazine advertisements, has been replaced by more focused marketing activities aimed at addressing individual customer needs. The advent of e-commerce activities and the wide range of commercial transactions which are automatically logged has dramatically accelerated the rate at which CRM analytics can be applied. Beside demographic and transactional data, the so-called online interaction data are achieving increasing relevance to CRM analysis. The dominant form of data here is internet clickstream data, although we must also include interactions that occur through wireless devices, cable television, e-mails, and more. Four main marketing tasks which have emerged within CRM: customer acquisition, customer retention, cross selling and up selling. All can be formulated as a classification problem. In the context of classification problems, one of the most successful approaches is represented by the theory of support vector machines (SVM) proposed by Vapnik (1995, 1998) and based on the structural risk minimization principle, which formally establishes the concept of reducing the empirical classification error as well as the generalization error, in order to achieve a higher accuracy on unseen data (Cristianini and Shawe-Taylor, 2000)(Schölkopf and Smola, 2002). Recently, the new concept of discrete support vector machines (DSVM) was introduced in different variants for binary classification problems (Orsenigo and Vercellis, 2004a, 2004b, 2004c; La Torre and Vercellis, 2003). The core idea of this approach is that the misclassification error evaluated over the training set, called empirical risk in the theory of SVM, be directly incorporated into the objective function, leading to mixed integer programming optimization problems, instead of resorting to a continuous proxy of the discrete error as for classical SVM. DSVM can be utilized for classification as linear perceptrons or framed within procedures for the construction of decision trees, deriving a multivariate discriminant function at each node. The computational testing on well-known benchmark datasets indicates that classifiers based on DSVM outperform other classification approaches in terms of accuracy. In particular, it is shown that even when the algorithm is forced to build trees with only one rule and two leaves, representing therefore linear perceptrons, yet it still achieves the highest accuracy among all competing techniques. The DSVM approach has been recently extended to multi-classification problems (Orsenigo and Vercellis, 2003b). The process of marketing campaign optimization takes a set of offers, a set of customer segments and a set of communication channels, and determines the most profitable combinations by which offers should go to segments over channels, taking into account a set of constraints for the campaign (Vercellis, 2002).

The sixth objective of the project concerns the distribution and the evaluation of public services (in fields as public health, environment, energy, transports). This is a field in which more and more sophisticated tools of quantitative analysis are required. In general, no preliminary analysis is done in order to obtain the maximum of the benefits and the minimum of the costs. In many cases only the profit is considered and other objectives, such as enviromental impact or social welfare, are not taken into account. The aim of the research in this context is to show how sophisticated tools of data mining can help institutions to take better decisions.
E-Government is defined as the use of information and communication technology in public administrations combined with organizational change and new skills in order to improve public services and democratic processes and strengthen support to public policies (Beckford 2001, Borroughs et al 2000, Bowerman 2002, DETR 1998, Klazinga 2000, McIver 1998, Packwood et al 1998, Pollitt 1996). In this respect, data mining tools can help e-government applications to be developed and be effective.
Service quality is a common topic of research in management and marketing science from the end of the 1970s (Williams, B. 1994, Cabinet Office 1999, Campbell, S. M. et al 2000, Freeman, T. 2002). The nature of service quality is different from the quality of physical goods and it is strongly related with traditional marketing concepts, like customer satisfaction and client loyalty. Since early 1980s researchers have made some models of service quality, based on these latter concepts. Our aim is to develop new models, based on data mining methodologies, and compare them with the previous models.

2.2.a Riferimenti bibliografici

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. (1995). Fast discovery of association rules. In: Advances in knowledge discovery and data mining. AAAI/MIT Press, Cambridge.
Baldi, P., Frasconi, P. e Smyth, P. (2003) Modelling the internet and the web. Wiley, New York.
Brooks, S., Giudici, P. and Roberts, G. (2003). Efficient construction of reversible jump MCMC proposal distributions. Journal of The Royal Statistical Society, series B, 65, pp 3-55, with discussion.
Chakrabarti, S. (2003) Mining the web: discovering knowledge from hypertext data. Morgan Kaufmann, New York.
Giudici, P. (2003), Applied Data Mining: statistical methods for business and industry, Wiley, London.
Giudici, P. (2001) Bayesian data mining, with application to credit scoring and benchmarking. Applied Stochastic Models in Business and Industry, 17, pp. 69-81.
Giudici, P. and Castelo, R. (2001) Association models for web mining. Knowledge discovery and data mining, 5, pp. 183-196.
Giudici,P. and Castelo, R. (2003) Improving MCMC model search for data mining. Machine learning, 50, pp 127-158
Giudici, P. and Cornalba, C. (2004) Statistical models for operational risk management. To appear in Physica A.
Giudici, P. and Green, P. (1999) Decomposable graphical gaussian model determination. Biometrika, 86, pp 785-801.
Giudici, P. and Passerone, G. (2001)Data Mining of association structures to model consumer behaviour. Computational Statistics and data analysis, 28, pp 533-541.
Han J., Kamber M. (2000), Data Mining: Concepts and Tecniques, Morgan Kaufmann.
Hand D. J., Heikki M., Smyth P. (2001), Principles of Data Mining, MIT Press.
Hastie, T., Tibshirani, R., Friedman, J. (2001), The elements of statistical learning: data mining, inference and prediction. Springer-Verlag.
Schoier, G. and Giudici, P. (2001). Cluster analysis for web usage mining. Italian Journal of Applied statistics,2.

Apostolico A. (2003). Pattern Discovery and the Algorithmics of Surprise (Invited Paper). In P. Frasconi, R. Shamir, EDS. Artificial Intelligence and Heuristic Methods for Bioinformatics. (pp. 111-127).
Apostolico A., M.E. Bock, and S. Lonardi. (2003). Monotony of Surprise and Large Scale Quest for Unusual Words. Journal of Computational Biology. vol. 10, pp. 283-311.
Apostolico A., M. J. Atallah. (2002). Compact Recognizers of Episode Sequences. Information and computation. vol. 15, pp. 481-494.
Apostolico A., M.E. Bock, S. Lonardi, X. Xu. (2000). Efficient Detection of Unusual Words. Journal of computational biology. vol. 7, pp.71-94.
A. Apostolico and Z. Galil (eds.), Pattern Matching Algorithms, Oxford Univ Press, 1997.
M. Crochemore and W. Rytter, Text Algorithms, Oxford University Press, New York, 1996.
R.O.Duda, P.E.Hart, D.G. Stork, Pattern Classification, Wiley, 2000.
E. Ferrari and G. Haus, The Musical Archive Information System at Teatro alla Scala. In Proc. of the IEEE International Conference on Multimedia Computing and Systems (ICMCS'99), pp. 817-821, Firenze, Italia, IEEE Computer Society Press, 1999.
Platt D., Guerra C., Rigoutsos I., Zanotti G. (2003). Globally induced interactions between secondary structures in proteins. Proteins-structure function and genetics. vol. 53, pp. 252-261.
Guerra C., S. Lonardi, G. Zanotti. (2002). Protein Structure Analysis using Indexing Techniques. 1st Int. Symposium on 3D Data Processing Visualization and Transmission. (pp. 812-821).

Deci, E. L. (1975). Intrinsic motivation. New York: Plenum Press.
Deci, E. L., e Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum Press.
Fasolo, B., Motta, M., e Misuraca, R. (2003). Il processo decisionale online: Rassegna di studi empirici e confronto tra Siti Internet per l'Aiuto alle Decisioni negli Stati Uniti e in Europa (in revisione).
Huber, J., Payne, J.W., e Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. Journal of Consumer Research, 9, pp. 90-98.
Iyengar, S. S., e Lepper, M. R. (2000). When Choice is Demotivating: Can One Desire Too Much of a Good Thing? Journal of Personality and Social Psychology, 76, 995-1006.
Payne, J.W., Bettman, J.R., Coupey, E., e Johnson, E.J. (1992). Multiple strategies in judgment and choice: a constructive process view of decision making. Acta Psychologica, 80, 107-141.
Slovic, P. (1995). The construction of preference. American Psychologist, 50, 364-371.
Taylor, S. E. (1989). Positive illusions: Creative self-deception and the healthy mind. New York: Basic Books.
Von Neumann, J., e Morgenstern, O. (1947). Theory of games and economic behavior (2nd ed.). Princeton NJ: Princeton University Press.

R. Spence, "Information Visualization", Addison Wesley/ACM Press, 2001
G. W. Furnas, "Effective View Navigation", ACM, Proceedings CHI '97
Learnframe, "Facts, Figures & Forces Behind e-Learning", August 2000 (rapporto on-line)
V. Cantoni, M. Cellario, M. Porta, "Perspectives and Challenges in E-Learning: Towards Natural Interaction Paradigms", Journal of Visual Languages and Computing, Elsevier Science Publishing, accepted July 2003

Berry, M. and Linoff, G. (1997). Data Mining techniques for marketing, sales, and customer support. Wiley, New York.
La Torre, D., Vercellis, C., 2003. C1,1 approximations of generalized support vector machines, Journal of Concrete and Applicable Mathematics 1, 125-134.
Normann, R., Ramirez, R. (1994) Designing interactive strategy. From value chain to value constellation, John Wiley & Sons Ltd.
Orsenigo, C., Vercellis, C., 2004a. Discrete support vector decision trees via tabu-search. Journal of Computational Statistics and Data Analysis, to appear.
Orsenigo, C., Vercellis, C., 2003a. Multivariate classification trees based on minimum features discrete support vector machines. IMA Journal of Management Mathematics vol. 14, pp. 221-234.
Orsenigo, C., Vercellis, C., 2004c. Rule induction through discrete support vector decision trees. In: Data Mining and Knowledge

Discovery Approaches Based on Rule Induction Techniques, E. Triantaphyllou & G. Felici eds., Kluwer Academic Publishers, Dordrecht, to appear.
Peppers, D., Rogers, M., 1993. The one to one future. Building relationships one customer at a time. Currency Doubleday, New York.
Schölkopf, B., Smola, A.J., 2002. Learning with kernels. Support vector machines, regularization, optimization and beyond. The MIT Press, Cambridge, USA.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York, USA.

Beckford, J. (2001) Quality, London: Routledge
Borroughs, T. E. et al (2000)Using root cause analysis to address patient satisfaction and other improvement opportunities, The Joint Commission Journal on Quality Improvement, 26(8):439-449
Bowerman, M. (2002) Isomorphism Without Legitimacy? The Case of the Business Excellence Model in Local Government, Public Money and Management, 22(2):47-51
Cabinet Office (1999) Involving users: Improving the delivery of local public services, London
Campbell, S. M. et al (2000) Defining quality of care, Social Science and Medicine, 51:1611-1625
DETR (1998) Modern Local Government: In Touch with the People, London
Freeman, T. (2002) Using performance indicators to improve health care quality in the public sector: a review of the literature, Health Services Management Research, 15, 126-137
Klazinga, N. (2000) Re-engineering trust: the adoption and adaptation of four models for external quality assurance of health care services in Western European health care systems, International Journal for Quality in Health Care, 12(3):183-189
McIver, S. (1998) Healthy Debate? An independent evaluation of citizens' juries in health settings, London: King's Fund
Packwood, T. et al (1998) Good Medicine? A case study of business process re-engineering in a hospital, Policy and Politics, 26(4): 401-415
Pollitt, C. (1996) Business approaches to quality improvement: why they are hard for the NHS to swallow, Quality in Health Care, 5:104-110

2.3 Numero di fasi del Programma di Ricerca:

6

2.4 Descrizione del Programma di Ricerca
Fase 1

Durata e costo previsto

Durata *Mesi 24* **Costo previsto** *Euro 61.000*

Descrizione

Testo italiano
Metodologie di web usage mining.

I ricercatori principalmente coinvolti saranno: Giudici, Cattaneo, Cerchiello, Dalla Valle, Figini, Bottinelli (Pavia); Schoier, Monte, Rondini, Melfi (Trieste); Cusano, Fini, Ferrari, La Torre, Salini, Termine, Tommasi (Milano)

La componente metodologica della ricerca farà prevalentemente riferimento allo sviluppo di modelli associativi per l'analisi di dati di web usage. Ciò verrà svolto seguendo le fasi tipiche del processo di data mining: obiettivi d'analisi, organizzazione dei dati, specificazione dei modelli statistici, elaborazione dei dati, confronto e valutazione dei modelli, interpetazione dei risultati. Dal punto di vista applicativo verranno analizzati dati reali di web usage mining e confrontati i risultati con quanto esistente in letteratura. Sottolineiamo da ultimo che particolare attenzione verrà dedicata a due tematiche di particolare interesse per il web usage mining: lo studio di metodologie per il pre-trattamento dei dati, soprattutto rivolte a problemi collegati al data fusion e lo lo sviluppo di procedure di Data Mining con riferimento spaziale.

Schematicamente le attività di ricerca previste sono le seguenti.
1a. Interpretazione statistica delle regole associative e sequenziali diffuse nella letteratura informatica. In particolare sviluppo e confronto di indici statistici che ne misurino la loro rilevanza a seconda dei diversi obiettivi di analisi (frequenza, capacità predittiva, interconnessione),
1b. Implementazione di modelli statistici basati sulle catene di Markov di primo ordine; sviluppo di modelli di ordine superiore e confronto con i precedenti. Sviluppo di modelli hidden Markov models, dove la variabile latente descrive comportamenti omogenei "di gruppo"
1c. Confronto fra i risultati ottenibili con l'applicazione delle regole associative e con l'impiego di modelli Markoviani.
1d. Considerazione degli aspetti di pre-processing, in particolare riguardanti data fusion e dati di natura spaziale
2: Confronto con quanto disponibile nel software SAS, ed implementazione delle eventuali componenti aggiuntive. Ad esempio per quanto riguarda i modelli Markoviani. Confronto con quanto disponibile nelle librerie del software R
3: Sviluppo di metodologie di valutazione dei modelli, di tipo computazionale, applicabili sia alle regole associative che ai modelli inferenziale di tipo Markoviano. Eventuale impiego di funzioni di perdita.

4: Esame della letteratura economico-aziendale esistente, e discussione del suo impatto sui modelli statistici sviluppati.
5: Sviluppo di modelli grafici bayesiani, in particolare mediante la determinazione di adeguati meccanismi di elicitazione della distribuzione iniziale
6. Dapprima applicazione a dati di log-files oggetto di precedenti analisi: uno proveniente dai percorsi di visita al sito Microsoft; l'altro riguardante un sito di e-commerce, i cui dati sono stati forniti da SAS. In seguito si attingeranno alle relazioni aziendali del laboratorio di data mining dell'Università di Pavia per ottenere nuove basi di dati.

Testo inglese
Web usage mining methodologies

The researchers mainly involved in this project will be: Giudici, Cattaneo, Cerchiello, Dalla Valle, Figini, Bottinelli (Pavia); Schoier, Monte, Rondini, Melfi (Trieste); Cusano, Fini, Ferrari, La Torre, Salini, Termine, Tommasi (Milano)

The methodological component of the research will mainly refer to the development of association models for web usage mining. This will be carried out following the typical data mining process phases: definition of the objectives of the analysis; organisation of the data; exploratory analysis; model specification; data processing; model comparison and assessment; interpretation of the results. We shall dedicate special attention to two important pre-processing problems arising in web usage mining: the treatment of data deriving from data fusion and of data having a spatial nature.
From the applied viewpoint we shall analyse real datasets taken from available logfiles, and compare the results with what available in the literature.
Schematically we plan the following research activities.
1a. Statistical interpretation of the association and sequence rules well known in the computer science literature. In particular development of statistical indexes that measure their relevance, according to the objectives of the analysis; and their subsequent interpretation (in tems of frequency, predictive power, interdependence)
1b. Implementation of statistical models based on first order Markov chains; development of higher order models; development of hidden Markov models, with a latent variable describing "clustering" behaviours
1c. Comparison of the results obtained with the classical association rules with those obtainable with Markov chains. Comparison with other graphical models.
1d. Consideration of pre-processing aspects, in particular involving data fusion and the analysis of spatial data
2: Comparison of what developed with what available in the softwares SAS and R. Possible writing of new routines to implement the proposed models
3: Development of appropriate model comparison methodologies, to compare methods in a fair way. Employment of loss functions.
4: Analysis of the business and marketing literature on the subject, and evaluation of its impact on statistical modelling.
5: Construction of Bayesian models, in particular graphical, and research on prior elicitation.
6. Applications: first to datasets already analysed (Microsoft and SAS case studies in Giudici, 2003); then to new datasets provided by companies in contact with the data mining laboratory of the University of Pavia.

Risultati parziali attesi

Testo italiano
-Costruzione di modelli di web usage mining compliant con gli standard di analisi e di processo disponibili in letteratura e nelle best practice internazionali;
- Pubblicazione di articoli scientifici in riviste e monografie, specialmente internazionali, relativamente ai modelli sviluppati;
- Sviluppo di algoritmi e di software conformi alla teoria proposta, di facile uso e sufficientemente scalabili;
- Pubblicazione di articoli divulgativi della metodologia, e delle sue potenzialità applicative, su riviste e monografie settoriali, al fine di contribuire al miglioramento delle prassi d'uso aziendali;
- Disseminazione dei risultati, in particolare mediante seminari e corsi di formazione, rivolti anche all'esterno, al fine di sensibilizzare la cultura generale alle esigenze emergenti da una attenta valutazione dei dati disponibili.

Testo inglese
- Construction of statistical models for web usage mining, compliant with the data mining process standards available in the literature as well in the best international practices;
- Publication of scientific papers, in journals and monographs, especially at the international level;
- Development of algorithms and software tools consistent with the proposed methods, possibly easy to use and sufficiently scalable;
- Publication of divulgative papers and reports in specialised reviews and books, with the aim of improving the cultural perception of data mining methods and of encouraging the adoption of practices based on such methods;
- dissemination of the results of the research, through seminars and courses directed outside of the network. In particular, organisation of workshops and conferences aimed at consolidating our network and to establish new research links.

Unità di Ricerca impegnate

Unità n. 2
Unità n. 4
Unità n. 5

Fase 2

Durata e costo previsto

Durata *Mesi 24* **Costo previsto** *Euro 32.000*

Descrizione

Testo italiano
Metodologie di web pattern discovery

I ricercatori che verranno prevalentemente coinvolti sono:
Apostolico, Ferrari, Guerra, Comin, Pizzi (Padova); Frasconi, Costa, Dubuisson, Ceroni (Pavia)

La componente metodologica farà riferimento allo sviluppo di modelli statistici per l'analisi di dati di web pattern discovery. Sulla base di tali modelli, saranno messi a punto algoritmi efficienti per trovare pattern con particolari caratteristiche di regolarita' ed anomalie e per scoprire episodi frequenti in una sequenza di eventi. Gli algoritmi utilizzeranno varianti, bilanciate per spazio-tempo, di costrutti esistenti quali alberi ed array di suffissi. Queste varianti della struttura dati dovranno essere costruite in tempo lineare sia a partire dall'albero dei suffissi che dall'array dei suffissi del testo dato. Il lavoro teorico che deve essere fatto consiste nella estensione di questi costrutti a modelli probabilistici meno semplicistici di quelli già trattati.

Schematicamente le sottofasi della ricerca previste sono le seguenti:

1: Analisi e sperimentazione con tecniche ed algoritmi esistenti per il web pattern discovery.
2: Estensione dei modelli esistenti e sviluppo di nuovi strumenti efficienti per l'analisi statistica di sequenze basati su modelli meno semplicistici di quelli gia' trattati
3: Sviluppo di metodi efficienti per la memorizzazione del numero di occorrenze in una sequenza di tutte le sue sottosequenze mediante speciali indici digitali di ricerca, alberi o automi
4: Sviluppo di tecniche di indicizzazione multidimensionale per accesso e ricerca dati di natura diversa (testi, immagini, video).
5. Da un punto di vista più applicativo, si procederà allo sviluppo di procedure per la localizzazione automatica e l'estrazione di informazioni di tipo specialistico dal Web e la loro messa a punto in un ambiente di calcolo distribuito. Lo scopo di tali sistemi è la generazione automatica di portali informativi specialistici. Esempi interessanti di sistemi simili per la classe di argomenti che comprende la letteratura scientifica informatica sono CiteSeer e CORA.

Testo inglese
Web pattern discovery methodologies

The researchers mainly involved in the project will be:
Apostolico, Ferrari, Guerra, Comin, Pizzi (Padova); Frasconi, Costa, Dubuisson, Ceroni (Pavia)

The methodological component of the research will insist on the development of statistical models for web pattern discovery. On the basis of such models, we shall develop efficient algorithms to find patterns with special regularities and anomalies and to discover frequent episods in a series of events. Such algorithms will employ modifications of existing constructs, such as trees and suffix arrays. These new proposals should be built in linear time, both from the suffix tree and from the suffix array of a given text. The thoretical work that need to be done consist in using more sophisticated underlying statistical models.

Schematically the research phases we plan to follow are:
1: Analysis and experimentation with esisting techniques and algorithms for web pattern discovery.
2: Extension of the existing models and development of new efficient tools for the statistical analysis of sequences based on models more complicated than those already used.
3: Development of efficient methods to memorise to number of occurrences in a sequence of all its subsequences by means of special digital research index, trees or robots.
4: Development of techniques of multidimensional indexing for the access and search of data of different nature (texts, images, video).
5. From a more applied viewpoint, we shall develop procedures for the automatic localization and extraction of specialized information from the Web and their implementation in a large distributed environment. The goal of such systems is the automatic generation of web portals with specialized content. Interesting examples of such systems for the class of subjects including the scientific literature in computer science are provided by CiteSeer e CORA.

Risultati parziali attesi

Testo italiano
- Sviluppo di software per la scoperta di pattern con particolari caratteristiche di regolarita' ed anomalie
- Pubblicazione di articoli scientifici in riviste e monografie, specialmente internazionali, sui modelli sviluppati
- Pubblicazione di articoli divulgativi della metodologia, e delle sue potenzialità applicative, su riviste e monografie settoriali, al fine di contribuire al miglioramento delle prassi aziendali
- Disseminazione dei risultati, in particolare mediante seminari e corsi di formazione, rivolti all'esterno, al fine di sensibilizzare la cultura alle esigenze emergenti da una attenta valutazione dei dati disponibili

Testo inglese
- Development of software tools and routines for web pattern discovery, dedicating special attention to patterns with anomalies and to non routinary behaviour;
- Publication of scientific papers, in journals and monographs, especially at the international level;
- Construction of web portals, based on pattern discovery, with specialised contents;
- Publication of divulgative papers and reports in specialised reviews and books, with the aim of improving the cultural perception of web pattern discovery and of encouraging the adoption of practices based on such methods;
- Dissemination of the results of the research, through seminars and courses directed outside of the network. In particular, organisation of workshops and conferences aimed at consolidating our network and at establishing new research links.

Unità di Ricerca impegnate

Unità n. 2
Unità n. 3

Fase 3

Durata e costo previsto

Durata *Mesi 24* **Costo previsto** *Euro 30.000*

Descrizione

Testo italiano
Aspetti cognitivi nell'interazione uomo-internet

I ricercatori che verrano prevalentemente convolti sono: Cardaci, Misuraca, Caci, D'Amico, Di Gesù, Ardizzone, Lenzitti (Palermo); Cantoni, Cellario, Porta, Semenza (Pavia)

La ricerca riguarderà l'analisi del modo in cui vengono opertate le scelte su internet, con specifico riferimento all'influenza di specifici fattori contestuali, quali l'effetto attraction. Si prevede di riscontrare un maggior numero di scelte, un minor tempo impiegato, minori livelli di difficoltà percepita e maggiori livelli di soddisfazione nelle condizioni non-conflittuali rispetto che nelle condizioni conflittuali, indipendentemente dal numero di opzioni presentate (poche vs molte). Ciò a conferma del fatto che le preferenze piuttosto che fisse e stabilite a priori, sono costruite di volta in volta in funzione del particolare modo in cui un problema decisionale è presentato e delle specifiche relazioni esistenti fra le opzioni di scelta.

Schematicamente le attività della ricerca previste sono le seguenti:

1. Campionamento. In questa fase si procederà alla costituzione di un gruppo di 120 partecipanti adulti all'oscuro degli scopi e del filone di ricerca in oggetto.

2. Strumenti e procedura. All'interno di un disegno sperimentale between-subjects, i partecipanti saranno assegnati a caso ad una delle seguenti quattro condizioni:

1. set di scelta conflittuale con poche opzioni (n=30);

2. set di scelta conflittuale con molte opzioni (n=30);

3. set di scelta non-conflittuale con poche opzioni (n=30);

4. set di scelta non-conflittuale con molte opzioni (n=30).

Nelle condizioni 1 e 2 ai soggetti verrà presentato un set di scelta composto rispettivamente da 6 e da 23 opzioni in cui nessuna domina le altre.

Nelle condizioni 3 e 4 invece ai soggetti verrà presentato un set di scelta composto rispettivamente da 6 e 23 opzioni in cui una (o oche) domina le altre.

Il compito dei soggetti consisterà nel compiere una scelta all'interno dello specifico set presentato. Verrà chiesto inoltre di compilare un breve questionario al fine di valutare, secondo una scala Likert a cinque livelli, la difficoltà percepita della decisione e il grado di soddisfazione legato alla scelta compiuta.

E' prevista una integrazione con l'unità di Pavia per quanto riguarda la messa a punto dei siti web sperimentali.

3. Analisi dei dati. In ciascuna condizione verrà rilevato il tempo impiegato dai soggetti per risolvere il compito. Il coefficiente rho di Spearman sarà utilizzato per calcolare la correlazione fra: numero di scelte compiute, tempo impiegato, grado di difficoltà percepita e livelli di soddisfazione.

Testo inglese
Cognitive aspects in man-internet interaction

The researchers that will be mainly dedicated to this research are: Cardaci, Misuraca, Caci, D'Amico, Di Gesù, Ardizzone, Lenzitti (Palermo); Cantoni, Cellario, Porta, Semenza (Pavia)

The third research project is aimed to study how users make decisions in the Internet, focusing on the influence of specific contextual factors, as the attraction effect.
We expect a higher number of choices, a lower time employed, lower levels of perceived difficulty and higher levels of satisfaction in the non-conflictual conditions rather than in the conflictual ones, and this independently by the number of options showed (few vs. many). These results are in accord to the position according to which preferences are dictated by the particular way in which a decisional task is presented and by the specific relations existing among the choice options.

The research will be scheduled in the following activities:

1. Sampling. 120 Italian subjects will take part in the research on a voluntary basis.
2. Materials and Procedure. In a between-subjects design, participants will be assigned randomly to one of the following four conditions:
1. conflictual choice condition with few alternatives (n=30);
2. conflictual choice condition with many alternatives (n=30);
3. non-conflictual choice condition with few alternatives (n=30);
4. non-conflictual choice condition with many alternatives (n=30);
In conditions 1 and 2 participants will be exposed to a choice set including respectively 6 and 23 options in which none is dominating.
In the conditions 3 and 4 participants will be exposed to a choice set including respectively 6 and 23 options in which one (or few) is dominating.

Participants will be asked to make a choice among the alternatives shown. Participants will then be asked to answer to a brief questionaire with the aim to evaluate, according to a five-level Likert scale, the perceived difficulty and the degree of satisfaction associated with the choice made.
In order to create the experimental web sites we will work in collaboration with the Pavia unit research.

3. Data Analysis. For each condition the time employed to make a decision will be measured. The Spearman's Rho test will be used to calculate the correlation among the number of choices made, the time employed, and the level of perceived difficulty and satisfaction.

Risultati parziali attesi

Testo italiano
- Pubblicazione di articoli scientifici in riviste e monografie, specialmente internazionali, sui modelli sviluppati
- Pubblicazione di articoli divulgativi della metodologia, e delle sue potenzialità applicative, su riviste e monografie settoriali, al fine di contribuire al miglioramento delle prassi sperimentali
- In congiunzione con i progetti applicativi: sviluppo di software che agevoli la scelta del visitatore. Ciò sia nel campo dell'e-learning che dell'e-government che della relazione con il cliente
- Disseminazione dei risultati, in particolare mediante seminari e corsi di formazione, rivolti all'esterno, al fine di sensibilizzare la cultura alle esigenze emergenti da una attenta valutazione degli aspetti cognitivi nell'interazione uomo-internet

Testo inglese
- Publication of scientific papers, in journals and monographs, especially at the international level;
- Construction of softwares and tools that simplify the decision making processes of the users. This in the application areas of e-learning, e-government and customer relationship management.
- Construction of web sites, optimised for man-internet interaaction, from the viewpoint of decision making;
- dissemination of the results of the research, through seminars and courses directed outside of the network. This especially within the cognitive sciences. In particular, organisation of workshops and conferences aimed at consolidating our network and to establish new research links.

Unità di Ricerca impegnate

Unità n. 1
Unità n. 2

Fase 4

Durata e costo previsto

Durata *Mesi 24* **Costo previsto** *Euro 35.000*

Descrizione

Testo italiano
Applicazione all'e-learning

I ricercatori che verranno prevalentemente coinvolti in questo progetto sono: Cantoni, Mosconi,Cellario, Porta, Semenza (Pavia);Apostolico, Ferrari, Guerra, Comin, Pizzi (Padova); Cardaci, Misuraca, Caci, D'Amico, Di Gesu', Ardizzone, Lanzitti (Palermo)

Si intende sviluppare paradigmi visuali e tecniche di interazione che rendano possibile un'analisi efficiente della grande quantità di informazioni associate ai LO, al fine di poter identificare quelli maggiormente rispondenti alle proprie esigenze.

Prevediamo le seguenti fasi di ricerca.
1. Saranno dapprima analizzate diverse metodologie di visualizzazione, che porteranno ad altrettanti prototipi (più o meno completi). Tramite sessioni di test su più utenti si identificheranno quindi le soluzioni più promettenti e ci si focalizzerà sulla progettazione del sistema finale in base alle indicazioni ottenute.
2. Per quanto riguarda il browsing delle caratteristiche dei singoli LO, si trarrà spunto dalle tecniche più comunemente sfruttate nell'ambito della cosiddetta "Information Presentation" (branca dell' "Information Visualization"), naturalmente opportunamente adattate alle esigenze del nostro campo di applicazione. Tra queste, di particolare interesse ci sembrano le metodologie basate sulla combinazione di "focus" e contesto, i quali, pur fornendo una rappresentazione dettagliata solo per il particolare elemento in esame, lasciano "intravedere" le caratteristiche degli elementi "vicini" (dove, nel nostro caso, la vicinanza si riferisce alla similarità dei valori dei metadati dei LO). Un esempio classico di rappresentazione di questo tipo si può ritrovare nella "visualizzazione bifocale", che si basa sull'analogia con la metafora di una striscia di carta continua che scorre lungo due rulli paralleli: la porzione centrale della striscia è quella con il contenuto informativo principale, mentre le restanti aree simulano (con un effetto prospettico tridimensionale) i nuovi dati in arrivo e i vecchi che si allontanano.
3. Maggiori sforzi richiederà sicuramente il progetto di metodi per il confronto diretto dei dati relativi a più LO. In questo caso si partirà da classiche tecniche di presentazione tabulare per arrivare a metodi più complessi, quali quelli che sfruttano rappresentazioni tridimensionali o implicano forme avanzate di esplorazione dinamica da parte dell'utente. Poiché i tipi di dati che dovranno essere confrontabili sono prevalentemente non numerici (ad esempio il 'titolo', il 'tipo di contenuto', ecc.), le tecniche scelte dovranno puntare maggiormente sulla caratterizzazione visiva dei vari elementi.
4. Data la quantità di informazioni che si potrebbero dover confrontare, è poco probabile una visualizzazione esaustiva contemporanea di tutti i dati di tutti i LO rispondenti a certe caratteristiche. Occorrerà quindi anche sviluppare tecniche di esplorazione dinamica (una sorta di "browsing" su gruppi di dati correlati) che permettano di muoversi all'interno di un universo comparativo relativamente ampio. Particolarmente utile potrà anche essere la possibilità di assegnare dinamicamente (tramite il sistema visuale) diversi "gradi di importanza" ai diversi metadati dei LO, così da modificare in tempo reale la visualizzazione in base alle specifiche esigenze.
5. Nella fase di sperimentazione, oltre al test dei prototipi preliminari a cui si è accennato, si sfrutteranno diversi repository di LO come fonte primaria di informazioni per la codifica visuale dei metadati. Saranno poi esplorate soluzioni didattico-decisionali avanzate nell'ambito del progetto MULTIMEDIA CAMPUS (Università di Pavia e Opera Multimedia), in simulazioni controllate di laboratorio e in sperimentazioni di progetti di formazione in corso (Master E.S.A.S. in Scienza e Tecnologia dei Media, Italia e Tunisia, presso l'Istituto di Studi Superiori, IUSS, dell'Università di Pavia) e in fase di avvio (Italia-Costarica).

Testo inglese
Application to e-learning

The researchers that will mainly be dedicated to this project are: Cantoni, Mosconi,Cellario, Porta, Semenza (Pavia);Apostolico, Ferrari, Guerra, Comin, Pizzi (Padova); Cardaci, Misuraca, Caci, D'Amico, Di Gesu', Ardizzone, Lanzitti (Palermo)

We intend to create visual paradigms and interaction techniques which make possible to efficiently analyze the great amount of data associated with Learning Objects (LOs), to identify those which better meet one's needs. We will exploit principles of Information Visualization, starting from well-known techniques to explore new solutions able to adapt to our specific application field. In particular, the visual systems which will be developed must allow great amount of information to be processed in parallel, through interaction paradigms close to human cognitive processes.

We plan the following research activities.
1. In a preliminary phase of the project, different visualization methodologies will be analyzed, which will lead to the development of some prototypes. Through test sessions involving several users, the best solutions will then be selected, and the focus will be shifted to the design of the final system according to the information obtained.
2. As regards feature browsing for the individual LO, techniques most commonly exploited within the so-called Information

Presentation (branch of Information Visualization) will be taken as a starting point. Among the possible strategies, we think that methodologies based on the combination of "focus" and context are especially interesting. A classical example can be found in the "bifocal display", which is based on the analogy with the metaphor of a continuous strip of paper running on two parallel cylinders: the central portion of the strip is the one containing the main information content, while the other areas simulate (with a perspective 3D affect) the new data which are coming and departing.
3. We will then proceed with the design of methods for direct comparison of data relative to several LOs. Also in this case, classical tabular visualization techniques will be exploited at first. Subsequently, we shall use more complex methods based on three-dimensional representations or involving advanced forms of dynamic exploration. Since the data types which must be comparable are mostly non-numerical (e.g. 'title', 'kind of content', etc.), the selected techniques will be mostly focused on a visual characterization of the various elements. Several solutions which could be adopted range from two-dimensional representations exploiting several visual codes (e.g. size, shape, color, position) to three-dimensional formalisms able to multiply the quantity of data which can be displayed at the same time.

4. Because of the great amount of information which must be potentially made comparable, a complete simultaneous visualization of all data relative to all LOs satisfying certain requirements is rather unlikely. Therefore, dynamic exploration techniques will be also developed (a kind of "browsing" on correlated data sets) which allow many data to be compared at once. Also, especially useful will be the possibility to dynamically assign (through the visual system) different "degrees of importance" to the various metadata of LOs, so as to modify, in real-time, the display modality according to specific requirements.

5. In an experimentation phase, besides the already quoted preliminar tests on early prototype systems, different LO repositories will be exploited as primary sources of information for visual coding of metadata. Then, advanced didactic solutions will be explored, within the project MULTIMEDIA CAMPUS (University of Pavia and Opera Multimedia), with simulations carried out in laboratory environments and with the experimentation within current training projects (Master in Media Science and Technology of the Institute of Advanced studies of the University of Pavia, both in Pavia and Tunis) and future ones (Italy-Costarica).

Risultati parziali attesi

Testo italiano
- Pubblicazione di articoli scientifici in riviste e monografie, specialmente internazionali, sui modelli sviluppati

- Pubblicazione di articoli divulgativi della metodologia, e delle sue potenzialità applicative, su riviste e monografie settoriali, al fine di contribuire al miglioramento delle prassi sperimentali

- Sviluppo di soluzioni di e-learning content management che agevolino la scelta del compositore.

- Sperimentazione delle soluzioni nell'ambito di progetti di spin-off universitari in corso (Multimedia Campus, società partecipata da Opera Multimedia e Università di Pavia)

- Disseminazione dei risultati, in particolare mediante seminari rivolti all'esterno.

Testo inglese
- Publication of scientific papers, in journals and monographs, especially at the international level, on the developed procedures;
- Development of e-learning content management platforms that facilitate the choice of the contents, in particular by the suppliers of teaching programmes;
- Publication of divulgative papers and reports in specialised reviews and books, with the aim of increasing the awareness on the potentialities and facilites of e-learning solutions;
- Experimentation of the proposed solutions, in the context of the existing academic spin-offs on the subject (Multimedia Campus, society jointly participated by the University of Pavia and the private company Opera multimedia)
- dissemination of the results of the research, through seminars and courses directed outside of the network.

Unità di Ricerca impegnate

Unità n. 1
Unità n. 2
Unità n. 3

Fase 5

Durata e costo previsto

Durata *Mesi 24* **Costo previsto** *Euro 40.000*

Descrizione

Testo italiano
Applicazione al customer relationship management

I ricercatori prevalentemente coinvolti in questo progetto saranno:
Vercellis, Orsenigo (Politecnico di Milano); Giudici, Cattaneo, Cerchiello, Figini, Dalla Valle, Bottinelli (Pavia); Schoier, Monte, Rondini e Melfi (Trieste)

Il progetto di ricerca riguarderà temi metodologici, con l'obiettivo generale di sviluppare e collaudare nuovi modelli e metodi matematici per data mining e knowledge discovery; in particolare, la ricerca si rivolgerà ai seguenti obiettivi specifici: metodi DSVM (discrete support vector machines) e SVM basati su kernels nonlineari; metodi DSVM e SVM per la classificazione multi-categorica; metodi DSVM e SVM per la classificazione di testi; sviluppo di criteri di misura e benchmarking per il confronto tra metodi di classificazione; metodi di selezione degli attributi in presenza di informazioni strutturate e semi-strutturate; metodi di classificazione di tipo PET (probability estimation trees) per l'assegnazione di punteggi individuali alle istanze di un dataset.

Si tratteranno inoltre temi applicativi legati al marketing relazionale, con l'obiettivo generale di sviluppare modelli, metodi e strumenti per l'ottimizzazione delle azioni di marketing; in particolare, la ricerca si rivolgerà ai seguenti obiettivi specifici: metodi rivolti all'analisi delle e-mail e delle clickstreams, e all'integrazione di queste informazioni con le più tradizionali informazioni demografiche e transazionali per sviluppare profilazioni e segmentazioni dei clienti; sviluppo di modelli di ottimizzazione multi-periodo per massimizzare la redemption e il ritorno sull'investimento delle campagne di marketing.

Schematicamente le attività della ricerca previste sono le seguenti:

Attività 1.1: Sviluppo di nuovi modelli e metodi matematici per specifici problemi di data mining. Questa attività è connotata da un alto profilo scientifico, e si rivolge allo sviluppo di modelli e metodologie matematiche che risultino innovativi nei rispettivi ambiti e che consentano di migliorare, in termini di efficienza nei tempi di elaborazione o di accuratezza nelle capacità di predizione, le prestazioni ottenute da altri metodi proposti in letteratura. L'attività di ricerca metodologica si rivolgerà a diversi specifici temi e problemi di data mining, che in generale abbiano attinenza e potenziale applicazione nell'ambito del marketing relazionale. Gli specifici temi investigati saranno quindi oggetto di diverse sottoattività.

Sottoattività 1.1.1: Sviluppo di metodi DSVM basati su kernels nonlineari per la classificazione binaria L'obiettivo di questa sottoattività riguarda l'estensione di kernels non-lineari alle discrete support vector machines, che già hanno dimostrato il loro potenziale predittivo rispetto ad altri metodi per problemi di classificazione binaria utilizzando kernels lineari.

Sottoattività 1.1.2: Sviluppo di metodi DSVM per la classificazione multi-categorica L'obiettivo di questa sottoattività riguarda l'estensione delle discrete support vector machines, con kernels sia lineari sia non-lineari, a problemi di classificazione multi-categorica. Già un primo esempio di estensione si è mostrato vantaggioso, mediante uno schema di adattamento di tipo one-versus-all. Si vogliono quindi esplorare altre forme più complesse di estensione, mediante schemi di tipo round-robin e schemi generali di tipo ECOC, nonché mediante la formulazione diretta di modelli di ottimizzazione intrinsecamente multi-categorici.

Sottoattività 1.1.3: Sviluppo di metodi DSVM di tipo PET per la classificazione L'obiettivo di questa sottoattività riguarda l'estensione di algoritmi basati su alberi di classificazione che utilizzino le DSVM per derivare i tagli in ogni nodo, alla classificazione probabilistica, al fine di assegnare un punteggio (score) a ciascuna istanza di un dataset. Tale estensione si rivela particolarmente significativa, oltre che sul piano teorico, anche per le numerose applicazioni nell'ambito del marketing relazionale, allorché sia necessario attribuire punteggi ai clienti per sviluppare una campagna promozionale rivolta alla conquista, alla retention, a cross e up selling.

Attività 1.2: Definizione di criteri e metriche per la valutazione e il confronto di tecniche di data mining Le tecniche di data mining rivolte alla classificazione sono innumerevoli, e divengono praticamente infinite al variare dei parametri che ciascuna di esse prevede al proprio interno. Si pone quindi con grande frequenza il problema di identificare, in uno specifico contesto predittivo, quale tecnica sia preferibile rispetto alle altre. Sul piano teorico, per identificare la tecnica più idonea in un contesto di marketing relazionale, è necessario sviluppare una metodologia rigorosa di confronto e benchmarking, che superi la diffusa consuetudine di limitarsi a un raffronto basato soltanto sulla percentuale di accuratezza.

Sottoattività 1.2.1: Identificazione dei criteri e delle metriche di confronto L'obiettivo di questa sottoattività riguarda l'identificazione di criteri obiettivi di confronto tra tecniche di classificazione, e la definizione di metriche che esprimano in termini quantitativi e misurabili i criteri individuati.

Sottoattività 1.2.2: Definizione di una metodologia di benchmarking L'obiettivo di questa sottoattività riguarda lo sviluppo di una metodologia di benchmarking, basata sulle metriche definite. In particolare, ci si rivolgerà a metodi basati sulla data envelopment analysis (DEA), già impiegata con successo dalla UR Polimi in altri ambiti applicativi per sviluppare confronto e benchmarking tra entità omogenee (in questo caso rappresentate da algoritmi alternativi).

Attività 1.3: Identificazione del contesto di riferimento per le applicazioni di data mining al marketing relazionale Questa attività ha l'obiettivo di identificare le principali esigenze delle imprese in rapporto all'applicazione di metodi analitici per il marketing relazionale, e il grado di diffusione di strumenti di business intelligence per analisi di marketing. Le informazioni verranno raccolte mediante indagini a campione.

Sottoattività 1.3.1: Identificazione degli ambiti di applicazione e delle informazioni da raccogliere. In questa prima sottoattività verranno identificati i principali settori di indagine (banche e assicurazioni, grande distribuzione organizzata, produttori di beni di largo consumo, etc.) e le informazioni da raccogliere nel corso delle interviste.

Sottoattività 1.3.2: Svolgimento di un'indagine empirica mediante surveys e case studies In questa sottoattività verranno svolte le indagini empiriche, sulla base di surveys basati su questionari inviati a un campione di imprese, integrato da case studies basati su interviste e indagini approfondite di alcune imprese particolarmente significative.

Sottoattività 1.3.3: Raccolta e analisi dei risultati In questa sottoattività si procederà all'analisi critica e all'interpretazione dei risultati ottenuti mediante le indagini empiriche, con l'obiettivo di tracciare lo stato di sviluppo della marketing intelligence nei settori considerati, e di identificare le principali esigenze evidenziate.

Attività 2.1: Sviluppo di nuovi modelli e metodi matematici per specifici problemi di data mining Anche nel corso della seconda fase proseguirà l'attività di ricerca a carattere metodologico, e si rivolgerà a problematiche di data mining legate all'analisi e alla classificazione di informazioni semistrutturate, che possono risultare rilevanti in applicazioni di marketing relazionale. L'attività di ricerca metodologica si rivolgerà quindi soprattutto al text mining e al web mining, nella prospettiva di integrare le diverse fonti di informazioni, demografiche, transazionali e quelle basate su visite on-line a siti e e-mail.

Sottoattività 2.1.1: metodologie per la classificazione di testi basati su DSVM e SVM L'obiettivo di questa sottoattività riguarda l'applicazione di modelli e tecniche di SVM e DSVM a problemi di classificazione automatica di testi, con particolare riferimento alla classificazione di e-mail.

Sottoattività 2.1.2: metodologie per la clickstream analysis basate su DSVM e SVM L'obiettivo di questa sottoattività riguarda l'applicazione di modelli e tecniche di SVM e DSVM a problemi di classificazione e predizione dei comportamenti di visitatori di siti di e-commerce, sulla base dell'analisi delle clickstreams, ovvero delle sequenze di interazioni e di pagine visitate.

Attività 2.2: Raccolta dei dati e costruzione di un archivio di dataset di benchmark L'obiettivo di questa sottoattività riguarda la costruzione di un archivio di dati benchmark per la validazione e il confronto di tecniche di data mining, con particolare attenzione ai problemi di classificazione, e con riferimento alle applicazioni di marketing relazionale. Si procederà alla raccolta di dati da problemi reali, all'eventuale perturbazione e filtraggio per evitare che questi evidenzino informazioni ritenute sensibili, alla catalogazione dei dataset per offrire accesso alle informazioni ad altri ricercatori.

Attività 2.3: Validazione e confronto di metodologie di data mining per il CRM. L'obiettivo di questa sottoattività riguarda la validazione e il confronto di metodologie alternative di data mining, con particolare riferimento alla classificazione, sulla base dei criteri e delle metriche definiti nel corso dell'attività 1.2, utilizzando i dataset di benchmark predisposti nel corso dell'attività 2.2.

Testo inglese
Application to customer relationship management

The researchers mainly involved will be:

Vercellis, Orsenigo (Politecnico di Milano); Giudici, Cattaneo, Cerchiello, Figini, Dalla Valle, Bottinelli (Pavia); Schoier, Monte, Rondini e Melfi (Trieste)

The planned research will deal with methodological themes, aiming at developing and testing new mathematical models and methods in the fields of data mining and knowledge discovery; in particular, the research will focus upon the following specific topics: DSVM (discrete support vector machines) and SVM methods based on nonlinear kernels; DSVM and SVM methods for multicategory classification; DSVM and SVM methods for text recognition; definition of measurement indexes to compare the performances among alternative classification approaches; feature selection and feature extraction methods for structured and semi-structured information; PET (probability estimation trees) to assign a score to each instance of a dataset.

The research will also consider applied themes, in the field of relational marketing, aiming at developing models, methods and tools supporting campaign management optimization; in particular, the research will focus upon the following specific topics; the analysis of clickstreams and e-mails and their subsequent integration with the more traditional demographic and transactional information in order to profile and segment the customer base; the development of new multi-period optimization models to maximize the overall return on the marketing investment.

The research conducted will be articulated through several activities as described below.

Activity 1.1: Development of new mathematical models and methods for solving specific data mining problems. This activity is characterized by a high scientific profile, and is aimed at developing new mathematical models and methods for different data mining problems. The new methods should improve, in terms of time complexity or prediction accuracy or both, the performances obtained by other techniques previously known and proposed in the literature. The activity of methodological research will address different specific themes and problems in the field of data mining, keeping in mind the general applicability to the relational marketing domain. The specific themes investigated will be framed within subactivities.

Subactivity 1.1.1: Development of DSVM methods based on nonlinear kernels for binary classification. This subactivity is aimed at extending the usage of nonlinear kernels to discrete support vector machines, which already have shown their high potential in prediction with respect to other techniques for binary classification, using linear kernels.

Subactivity 1.1.2: Development of DSVM methods for multi-categorical classification. This subactivity is aimed at extending the discrete support vector machines, both with linear and nonlinear kernels, to multi-categorical classification. A first result has showed the advantages of using DSVM for multi-categorical classification, by means of a one-versus-all scheme. Next steps will be oriented to exploit more complex forms of extension, such as round-robin and ECOC schemes. A direct formulation of the multi-classification problem as a large discrete optimization problem will be also investigated.

Subactivity 1.1.3: Development of DSVM methods as PET classifiers. This subactivity is aimed at extending classification algorithms based upon decision trees to probabilistic classification, making use of DSVM models to derive the partitions at each node. In this way a distinct score is assigned to each instance in a dataset. This extension is particularly relevant from the theoretical point of view, and also because of the applicability in the domain of relational marketing. In this case scoring is required to efficiently target a marketing campaign, aimed at obtaining new customers, or keeping loyal existing ones, or to cross and up sell products and services.

Activity 1.2: Identification of criteria and metrics for evaluating and comparing data mining techniques. There is a large number of alternative data mining techniques aimed at classification, and they become practically infinite in number by letting vary their parameters. Therefore, the problem arises of identifying, in a specific application domain, which method is more suited and should be preferred to other competing techniques. Hence, to properly deal with this issue, a theoretical framework should be developed to systematically and rigorously benchmarking alternative techniques. Simply resorting to a comparison based upon total accuracy level on a test dataset, which is often adopted in practice, is far from satisfactory, particularly when dealing with complex and realistic problems in the relational marketing domain.

Subactivity 1.2.1: Identification of criteria and metrics for comparison. This subactivity is aimed at identifying rigorous and meaningful criteria for comparing alternative classification techniques. These criteria will then be translated into metrics expressing in quantitative and measurable terms the criteria identified.

Subactivity 1.2.2: Definition of a benchmarking methodology. This subactivity is aimed at developing a benchmarking methodology, based on the metrics previously defined. In particular, methods based upon data envelopment analysis (DEA) will be considered. These methods have been successfully utilized by RU Polimi in other application domains, to derive comparisons and to benchmark homogeneous entities (represented in this case by alternative algorithms).

Activity 1.3: Identification of the application domain for data mining analysis in relational marketing. This activity is aimed at identifying the main needs that companies face in developing relational marketing strategies. The degree of dissemination of business intelligence tools and analytical methods for relational marketing will also be monitored. The information will be gathered through surveys on a sample of companies.

Subactivity 1.3.1: Identification of the application domains and the information to be collected. This subactivity is aimed at identifying the main industries involved in the surveys (banking and insurance, retail and distribution, producers of packaged consumer goods, etc.). The information to be collected during interviews will also be identified.

Subactivity 1.3.2: Conducting the empirical study by means of surveys and case studies. This subactivity is aimed at conducting the empirical studies, based on survey questionnaires addressed to a sample of companies. More detailed case studies will also be collected, dealing with companies of particular relevance.

Subactivity 1.3.3: Collection and analysis of the survey results. This subactivity is aimed at critically analyzing the results obtained through the empirical studies. This will allow to depict the state of the art with respect to the dissemination and usage of business intelligence in the industries investigated. The main needs of the companies will also be pointed out.

Activity 2.1: Development of new mathematical models and methods for solving specific data mining problems. This activitiy will be addressed to data mining problems arising in connection to the classification of semistructured information, relevant in the relational marketing domain. Research activities will therefore be oriented to text mining and web mining, with the aim of integrating different sources of information, such as demographic, transactional, on-line internet based, e-mail based.

Subactivity 2.1.1: Methodologies for text classification based on DSVM and SVM. This subactivity is aimed at developing models and techniques for text classification based on SVM and DSVM. In particular, the classification of e-mails will be addressed.

Subactivity 2.1.2: Methodologies for clickstream analysis based on DSVM and SVM. This subactivity is aimed at developing models and techniques based on SVM and DSVM for classifying and predicting the behavior of visitors of web sites, with particular emphasis on e-commerce sites. In particular, clickstreams will be analyzed, considering the sequence of pages visited and interactions with the users.

Activity 2.2: Data collection and creation of a benchmarking datasets repository. This activity is aimed at creating a benchmarking datasets repository to evaluate and compare alternative data mining techniques, focusing particularly on classification problems and relational marketing applications. This activity includes the collection of data from real world problems, data cleaning and filtering to remove noisy information, datasets cataloging to offer an easy information access to other researchers.

Activity 2.3: Assessment and benchmarking of data mining methodologies for CRM. This activity is aimed at evaluating and comparing alternative data mining techniques, focusing particularly on classification, on the basis of the criteria and metrics identified in activity 1.2 and using the benchmarking datasets created in activity 2.2.

Risultati parziali attesi

Testo italiano
-Costruzione di modelli di classificazione e di customer relationship management compliant con gli standard di analisi e di processo disponibili in letteratura e nelle best practice internazionali;
- Pubblicazione di articoli scientifici in riviste e monografie, specialmente internazionali, relativamente ai modelli sviluppati;
- Sviluppo di algoritmi e di software conformi alla teoria proposta, di facile uso e sufficientemente scalabile;
- Pubblicazione di articoli divulgativi della metodologia, e delle sue potenzialità applicative, su riviste e monografie settoriali, al fine di contribuire al miglioramento delle prassi d'uso aziendali;
- Disseminazione dei risultati, in particolare mediante seminari, corsi di formazione e convegni specialistici, rivolti anche all'esterno, al fine di sensibilizzare la cultura generale alle esigenze emergenti da una attenta valutazione dei dati disponibili.

Testo inglese
- Construction of classification models for customer relationship management, in particular making use of web usage mining data;
- Development of software tools and routines for customer relationship management, easy to use and scalable;
- Publication of scientific papers, in journals and monographs, especially at the international level;
- Publication of divulgative papers and reports in specialised reviews and books, especially concerned with direct marketing and customer relationship management, with the aim of improving the current CRM practices;
- dissemination of the results of the research, through seminars and courses directed outside of the network. In particular, organisation of workshops and conferences aimed at consolidating our network and to establish new research links.

Unità di Ricerca impegnate

Unità n. 2
Unità n. 5
Unità n. 6

Fase 6

Durata e costo previsto

Durata *Mesi 24* **Costo previsto** *Euro 48.900*

Descrizione

Testo italiano
Applicazione all'e-government

I ricercatori prevalentemente impegnati saranno: Cusano, Fini, Ferrari, La Torre, Salini, Termine, Tommasi (Università di Milano); Vercellis, Orsenigo (Politecnico di Milano); Giudici, Cattaneo, Figini, Bottinelli, Dalla Valle, Cerchiello (Pavia)

Due sono le aree su cui intendiamo lavorare:
- l'efficienza e l'efficacia dei servizi pubblici. L'obiettivo è di migliorare l'efficacia interna dei servizi e le interazioni tra diversi livelli e funzioni. A questo fine il gruppo di ricerca si concentrerà sulla riorganizzazione dei processi produttivi e delle risorse umane al fine di sviluppare sistemi per la condivisione delle informazioni.
- la qualità dei servizi pubblici. Il gruppo di ricerca punta a migliorare la qualità dei servizi pubblici attraverso la condivisione dei dati. La misura della qualità di un servizio è un problema cruciale. Ci sono tre tipi di misure della qualità di un servizio:
a) qualità tecnica e professionale, che è misurata attraverso variabili osservabili in accordo con le competenze e le conoscenze professionali.
b) qualità dell'erogatore, misurata attraverso un'analisi costi e benefici.
c) qualità del cliente, ovvero la qualità percepita, che è misurata tramite la customer satisfaction analysis.
La ricerca può essere divisa nelle seguenti fasi:
- analisi ambientale, benessere sociale
- valutazioni della domanda e delle performance del servizio
- modellistica e analisi degli obiettivi
- stima dei tempi
- valutazione dei rischi
- simulazione e previsione di scenari

Le fasi di lavoro possono essere schematizzate come segue:
1. Raccolta e organizzazione dei dati. Costruzione dei data warehouse. Per poter studiare le caratteristiche socio demografiche delle unità territoriali occorre disporre di informazioni articolate e complesse in generale racchiuse in datawarehouse e data mart

di grandi dimensioni. Il data mining è un processo interattivo e iterativo che deve essere usato congiuntamente con tecnologie avanzate per identificare relazioni nascoste e regolarità nei dati. I dati storici sono usati per generare modelli che possono essere applicati successivamente ad aree come la predizione, la previsione, la stima e il supporto alle decisioni.
2. Analisi esplorative preliminari dei dati. Modelli statistici di classificazione e segmentazione. Per ognuno dei gruppi individuati è possibile calcolare degli indicatori sintetici che forniscano un quadro della complessità della situazione esistente. Un'analisi esplorativa dettagliata delle unità territoriali e delle variabili che le caratteristizzano tramite le tecniche di analisi statistica classiche univariate, bivariate e multidimensionali (studio delle distribuzioni, connessione, correlazione, analisi della varianza, analisi dei cluster, analisi discriminante, analisi delle componenti lineare e non lineari) unitamente a modelli di scoring e di classificazione permettono di determinare, ad esempio, indicatori demografici e sociali, criteri di performance dei servizi erogati, legami tra servizi diversi e benchmarking fra erogatori.
3. Analisi multicriterio e decisioni. I servizi pubblici richiedono l'analisi congiunta di una molteplicità di obiettivi spesso concorrenti e contrastanti. I modelli multiobiettivo e l'analisi multicriterio facilitano la considerazione sulla stima dell'investimento che in molti casi non possono essere ridotti all'analisi economica come, per esempio, uguaglianza sociale, la protezione ambientale, uguaglianza di trattamento. I modelli multicriterio cercano di sintetizzare questa pluralità e permettono di determinare soluzioni efficienti dei problemi decisionali.
I problemi multicriterio sono spesso studiati attraverso tecniche di scalarizzazione che riducono il problema ad un solo obiettivo. Comunque questa procedura ha alcuni inconvenienti; solo alcune soluzioni efficienti del problema vengono determinate e, in molti casi, problemi regolari vengono trasformati in problemi irregolari. Molti problemi cercano di investigare soluzioni di problemi multicriterio senza procedure scalarizzazione.
4. Test di adattamento e valutazioni delle performance del modello. Analisi quantitativa e qualitativa dei modelli ottenuti con previsione di scenari e valutazione dei costi e benefici.

Testo inglese
Application to e-government

The researchers mainly involved are:
Cusano, Fini, Ferrari, La Torre, Salini, Termine, Tommasi (Milano); Vercellis, Orsenigo (Polytechnic of Milan); Giudici, Cattaneo, Figini, Bottinelli, Dalla Valle, Cerchiello (Pavia)

Two main areas will be addressed:
- the effectiveness and efficiency of public services. The goal is to improve the internal effectiveness of public services and the interaction between different levels and different functions. To this end the research group should address the need for re-structuring the working processes of public services: the archiving and warehousing of information collected and the development of systems to enable sharing of information.
- The quality of public services. The research group aims to improve the quality of public services through sharing data among them. The measurement of service quality is a core problem. There are three aspects and three kinds of measurement of service quality:
a) professional or technical quality, which is measured by observable variables according to professional knowledge and competence.
b) distributor quality, which is measured by cost-benefit analysis.
c) customer or client quality, which is measured by customer's evaluation of service.
Our research can be divided into the following steps:
- Enviromental analysis, social welfare.
- Analysis of demand and performance.
- Analysis of objectives and goals.
- Time analysis
- Risk evaluation
- Simulation and scenario forecasting.

The research activities for this objective can be summarised as follows:
1. Data collection and data warehousing. In order to study social and demografical characteristics it is necessary to have many information which are usually stored in data warehouse and data marts of big dimensions. Data Mining is an interactive and iterative process which must be used jointly with advanced technologies to identify underlying relationships and features in the data. Historical data are used to generate models, which can be applied at a later date to areas such as prediction, forecasting, estimation and decision support.
2. Preliminary analysis. Statistical models of classification and segmentation. For each group it is possible to build some indicators which describe the complexity of the system. A detailed analysis of units and of variables by classical multidimensional statistical techniques (connection, correlation, analysis of variance, cluster analysis, discriminant analysis, analysis of linear and non linear components) and classification models allow to calculate social and demographical indicators, performance criteria and relations among different services.
3. Multicriteria analysis. Public services requires a joint analysis of many objectives. Multiobjective analysis facilitates consideration in the investment appraisal of policy maker's objectives that in many cases can not be reduced to economic analysis (for example, social equity, enviromental protection and equal opportunities). Multicriteria models can describe a variety of objectives and give efficient solutions to decision problems. Multicriteria problems are often studied by scalarization techniques which reduce a multiobjective problem to a single one. However this procedures have some drawbacks; they give only some efficient solution of the initial problem and, in many cases, they reduce a smooth problem to a nonsmooth one. Many algorithms try to investigate the solutions of multiobjective problems without scalarizations.
4. Tests and evaluation of the model. Qualitative and quantitative analyses of the models will be used in order to forecast future situations and for the evaluation of costs and benefits.

Risultati parziali attesi

Testo italiano
-Costruzione di modelli di data mining per l'e-government compliant con gli standard di analisi e di processo disponibili in letteratura e nelle best practices;
- Pubblicazione di articoli scientifici in riviste e monografie, specialmente internazionali, relativamente ai modelli sviluppati;
- Sviluppo di algoritmi e di software conformi alla teoria proposta, di facile uso e sufficientemente scalabile;
- Pubblicazione di articoli divulgativi della metodologia, e delle sue potenzialità applicative, su riviste e monografie settoriali, al fine di contribuire al miglioramento delle prassi d'uso;
- Disseminazione dei risultati, in particolare mediante seminari, corsi di formazione e convegni specialistici, rivolti anche all'esterno, al fine di sensibilizzare la cultura delle istituzioni pubbliche alle esigenze emergenti da una attenta valutazione dei dati disponibili.

Testo inglese
- Construction of data mining models for e-government problems, compliant with the process standards available in the literature as well in the best institutional practices;
- Publication of scientific papers, in journals and monographs, especially at the international level;
- Development of algorithms and software tools consistent with the proposed methods, possibly easy to use and sufficiently scalable;
- Publication of divulgative papers and reports in specialised reviews and books, with the aim of improving the cultural perception of data mining methods and of encouraging the adoption of practices based on such methods;
- dissemination of the results of the research, through seminars and courses directed outside of the network.

Unità di Ricerca impegnate

Unità n. 2
Unità n. 4
Unità n. 6

2.5 Criteri suggeriti per la valutazione globale e delle singole fasi

Testo italiano
Dal punto di vista delle modalità di valutazione, evidenziamo che l'attività di ricerca del gruppo si avvarrà dei numerosi contatti ed esperienze di relazione gia' sviluppati dalle singole unità di ricerca, a livello scientifico, didattico ed applicativo.

Dal punto di vista scientifico verranno valorizzate le appartenenze di diversi membri del gruppo ad associazioni e gruppi di ricerca settoriali, in particolare riguardanti il data mining e l'e-business. Attualmente tali appartenenze sono su elevati standard qualitativi e coprono un'ampia gamma di settori; una loro messa in rete non può che produrre delle sinergie positive. Proponiamo che, nella valutazione globale della rete, venga sentito un panel di esperti internazionali appartenenti a tali associazioni e gruppi di ricerca, al fine di valutare il contributo del gruppo di ricerca allo svilupppo della ricerca del settore.

Dal punto di vista didattico, molti membri del gruppo di ricerca partecipano ad esperienze didattiche di eccellenza, a livello nazionale ed internazionale, nell'ambito del data mining e dell'e-business. La messa in rete di tali competenze favorirà il miglioramento degli standard didattici ed il loro aggiornamento. Poichè tali percorsi didattici comprendono lauree specialistiche, master e dottorati, proponiamo che, nella valutazione globale della rete, venga sentito un panel di studenti, coinvolti in tali corsi, al fine di valutare il contributo del gruppo di ricerca allo svilupppo delle conoscenze in ambito di data mining ed e-business

Infine, dal punto di vista applicativo, diverse unità di ricerca sono responsabili di un laboratorio di data mining o inerente le applicazioni di e-business. In tale ambito, sono già disponibili numerosi contatti con società di software e di servizi alle imprese, con società finanziarie ed industriali e con enti pubblici. Questi contatti consentiranno di tenere ben presenti nel corso della ricerca le effettive esigenze applicative delle imprese e degli utilizzatori di e-business. Inoltre potranno agevolare lo sviluppo della ricerca stessa, specie a livello applicativo. La messa in rete di tali contatti agirà come un forte moltiplicatore delle relazioni e permetterà di potenziare considerevolmente l'attività di laboratorio in essere. Proponiamo che, nella valutazione globale della rete, venga sentito un panel di aziende ed enti pubblici, rappresentativi della realtà nazionale, al fine di valutare il contributo del gruppo di ricerca allo svilupppo delle pratiche applicative in tema di data mining ed e-business.

Dal punto di vista dei criteri suggeriti di valutazione, suggeriamo che la ricerca del gruppo venga valutata seguendo otto indicatori principali, descritti di seguito.

1. Pubblicazione dei risultati della ricerca su riviste scientifiche di consolidata reputazione internazionale

2. Pesentazione dei risultati della ricerca a workshop e convegni, preferibilmente tematici, specialmente con relazioni invitate

3. Organizzazione, specie in collaborazione fra le varie unità di ricerca, di workshop di ricerca e convegni sul tema oggetto della

ricerca

4. Organizzazione, specie in collaborazione fra le varie unità di ricerca, di scuole di formazione per dottorandi, post-dottorandi e assegnisti di ricerca, sui temi oggetto della ricerca

5. Realizzazione e sperimentazione di soluzioni informatiche (software, routine, siti e portali internet), specialmente per gli aspetti più applicativi della ricerca

6. Collaborazioni con isituzioni ed imprese, in diverse forme, comprese attività di networking, di formazione e di ricerca in comune. Ciò con particolare riferimento all'istituzione di borse e contratti di ricerca finanziati dall'esterno.

7. Collaborazione e interazione scientifica fra le unità del progetto, sia consolidando esistenti relazioni di ricerca che attivandone di nuove, specie a livello interdisciplinare

8. Costituzione di una rete di ricerca attiva, capace di rappresentare la ricerca nazionale nel campo del web mining e, più in generale, del data mining. Grado di riconoscimento di un tale network dall'esterno, in particolare da network internazionali già esistenti.

Testo inglese
In terms of the actual modalities of evaluation, we underline that the research activity of the group will draw upon several contacts and experiences already developed and established at the individual research units. This from different aspects: scientific, didactic, applied.

From a scientific viewpoint we shall value the memberships of several researchers in international research groups and associations, in particular concerning data mining and e-business.
Often such memberships concern higly respectd networks and cover a wide range of fields; networking them will bring positive sinergies. We propose that, in the global evaluation of the research, a panel of international experts in the field, belonging to the above mentioned research groups, will be interviewed, with the aim of evaluating the actual contribution of our group to the the international research in the field.

From a didactic viewpoint many researchers in the group participate in training programmes, university degrees (including international Master's and Phds) and schools of excellence, in the fields of data mining and e-business.
The networking and sharing of such experiences will help improving teaching quality and updating of the programs. We propose that, in the global evaluation of the research, a panel of students of such courses will be interviewed, with the aim of evaluating the actual contribution of our group to the education in data mining and e-business.

Finally, from an applied viewpoint, several research units are responsible of laboratories, in the field of data mining or concerning specific e-business applications. In this framework, several contacts and joint activities with companies and instituions are already present. They concern a wide range of institutions, from the public sector, to IT and software vendors, financial and industrial companies, and consulting companies. These contacts will help the research to take into account real applied e-business problems. Furthermore, the sharing of such contacts wil help the group to become a credible and strong reference network in the research fields. We propose that, in the global evaluation of the research, a panel of company experts in the research fields will be interviewed, with the aim of evaluating the actual contribution to the actual data mining and e-business practices.

In terms of the suggested evaluation criteria, we propose that the research of the group will be evaluated according to eight main indicators, described below.
1. Publication of the results of the research in authoritative international scientific journals
2. Dissemination of the results of the research at conferences and workshops, preferably thematic, especially with invited talks
3. Organisation, especially in collaboration among the research units, of research workshop and conferences on the research themes
4. Organisation, especially in collaboration among the research units, of training courses and schools for Phd students, post-docs and junior researchers
5. Realisation, implementation and testing of Information technology solutions (softwares, routines, web sites and portals), especially for the applied aspects of the research
6. Collaborations with institutions and companies, at different levels, including the development of joint activities of networking, training, and research. This with special reference to the activation of scolarships and research contract, funded by outside partners.
7. Scientific collaborations and cross-fertilisations among the research units of the projects, both consolidating existing collaborations, and in establishing new ones (especially in a cross-disciplinary way)
8. Constitution of a research network, able to represent the italian research in the field of web mining and, more generally, data mining. Recognition of the status of such a network by existing intrnational networks.

3.1 Spese delle Unità di Ricerca

Unità di Ricerca	Voce di spesa										TOTALE
	Materiale inventariabile	Grandi Attrezzature	Materiale di consumo e funzionamento	Spese per calcolo ed elaborazione dati	Personale a contratto	Servizi esterni	Missioni	Partecipazione / Organizzazione convegni	Pubblicazioni	Altro	
Unità nº 1	*9.000*	*0*	*1.500*	*0*	*5.000*	*0*	*9.000*	*4.000*	*1.500*	*0*	30.000
Unità nº 2	*6.000*	*0*	*3.000*	*0*	*36.000*	*0*	*16.000*	*9.000*	*0*	*0*	70.000
Unità nº 3	*3.000*	*0*	*1.000*	*1.000*	*0*	*0*	*15.000*	*11.000*	*1.000*	*0*	32.000
Unità nº 4	*0*	*0*	*0*	*1.000*	*12.900*	*0*	*20.000*	*15.000*	*0*	*0*	48.900
Unità nº 5	*8.000*	*0*	*1.000*	*2.000*	*0*	*0*	*8.000*	*5.000*	*2.000*	*0*	26.000
Unità nº 6	*0*	*0*	*0*	*0*	*20.000*	*0*	*5.000*	*15.000*	*0*	*0*	40.000
TOTALE	**26.000**	**0**	**6.500**	**4.000**	**73.900**	**0**	**73.000**	**59.000**	**4.500**	**0**	**246.900**

3.2 Costo complessivo del Programma di Ricerca

Unità di Ricerca	Voce di spesa					
	RD	RA	RD+RA	Cofinanziamento di altre amministrazioni	Cofinanziamento richiesto al MIUR	Costo totale del programma
Unità n. 1	*0*	*9.000*	*9.000*	*0*	*21.000*	*30.000*
Unità n. 2	*6.000*	*15.000*	*21.000*	*0*	*49.000*	*70.000*
Unità n. 3	*7.200*	*2.400*	*9.600*	*0*	*22.400*	*32.000*
Unità n. 4	*0*	*14.700*	*14.700*	*0*	*34.200*	*48.900*
Unità n. 5	*0*	*8.000*	*8.000*	*0*	*18.000*	*26.000*
Unità n. 6	*12.000*	*0*	*12.000*	*0*	*28.000*	*40.000*
TOTALE	**25.200**	**49.100**	**74.300**	**0**	**172.600**	**246.900**

	Euro
Costo complessivo del Programma	*246.900*
Fondi disponibili (RD)	*25.200*
Fondi acquisibili (RA)	*49.100*
Cofinanziamento di altre amministrazioni	*0*
Cofinanziamento richiesto al MIUR	*172.600*

(per la copia da depositare presso l'Ateneo e per l'assenso alla diffusione via Internet delle informazioni riguardanti i programmi finanziati e la loro elaborazione necessaria alle valutazioni; legge del 31.12.96 n° 675 sulla "Tutela dei dati personali")

Firma _____ Data 31/03/2004 ore 17:34