Summary of the thesis

DATA INTEGRATION AND AUTOMATIC TEXT SUMMARIZATION: A PATH TO MORE INFORMED BUSINESS DECISIONS

In recent years, there has been an explosion of data shared online. The majority of this internet information is in text format, and can be used as a source to create new knowledge. These data are frequently unstructured and, in their raw state, cannot be used for any type of analyses, resulting challenging to manage from an Information Technology (IT) perspective. But in addition to these types of data, most companies have a huge collection of structured data, acquired and built over time. The union of these two types of information represents therefore a gold mine to be able to draw as much knowledge as possible from them. Because of this, the so-called Data Pre-processing (DPP), an important stage in the Data Mining process, allows significant manipulations on them, in order to make them useable for any subsequent elaboration procedure. The general DPP steps are the Data Cleansing, Data Integration, Data Reduction, and Data Transformation, while guaranteeing the protection of the privacy. This research focused on two different applications related to structured and unstructured data through respectively a focus on a Data Integration (DI) challenge, and one on the Automatic Text Summarization (ATS) task, whose algorithm evaluation metrics were explored. One of the most challenging issues in DI, is the research for automatic or semi-automatic methodologies, since these techniques often require the expertise of a domain specialist who can direct the process and improve the results. However, in the literature, there are not many fully or semi-automatic DI approaches unless they include experts with specific IT-skills. So, in this study, by the assistance of an intermediary figure (the *Company Manager*), who is not necessary skilled in IT, using an Information Retrieval methodology, clustering methods, and a trained neural network, we have built a semi-automatic DI process. This process is capable of reducing persistent conflicts in data, and ensuring a unified view of them, respecting the original constraints of the datasets and guaranteeing a high-quality outcome for Business Intelligence evaluations. At the same time, having the ability to reduce the amount of text from which to extract information is essential, when there are textual data sources involved. This is important to recover the key concepts, but also to speed up the analysis systems. In particular, ATS is a interesting challenge of Natural Language Processing. The primary issue is that there are currently a number of algorithms that attempt to reduce documents, using both statistical techniques (Extractive algorithms) and Artificial Intelligence methods (Abstractive algorithms). However, several metrics primarily based on the overlap analysis of *n-grams* such as the ROUGE, which is the most used, are applied to assess the quality of the results. Therefore, determining if these metrics are efficient, and

whether they really enable to compare the quality of the outcomes of the various Text Summarization (TS) algorithms, is the focus of the second research topic.

Chapter 1 presents the primary context of this research work, the Data Management. The initial overview considers the various challenges that need to be addressed when beginning any Data Analysis process starting from structured and unstructured data. The various steps that make up the Data Pre-processing phase are also described, an essential phase for the correct and efficient execution of any data analysis job.

Chapter 2 introduces all the Machine Learning (ML) and Deep Learning (DL) techniques used in this research work. In particular, it is focused on the description of hierarchical and non-hierarchical clustering techniques, analyzing their differences. Subsequently, neural network approaches are highlighted, making an overview of the various types considered in this study and finally, it shows the most important aspects related to the Information Retrieval (IR), the process of finding relevant information from a collection of data.

Chapter 3 is related to the DI challenge. Following a brief presentation of the fundamental concepts of this particular topic, the State of the art is carefully explored, through the investigation of various DI methodologies. Next, our new DI approach that merges distinct heterogeneous data sources using a semi-automatic procedure is described. Using an IR technique, Clustering methods and a Neural network, it is possible to conclude the process involving not an IT-expert, but a figure who will only act as a link between system developers and end users, who does not need to have IT skills to complete the task. In the integrated database, the outcomes will respect all the constraints between attributes existing prior to integration.

In *Chapter 4*, the topic of ATS algorithms is investigated. These techniques try to automatically extract important information from one or more input texts, creating summaries while retaining the meaning of the content. Following a brief presentation of the basic concepts related to text representation and text similarity, the State of the art of the ATS algorithms and their evaluation metrics in the literature is explored. The quality of the summaries produced by these algorithms has been evaluated using a variety of metrics, the most popular of which is ROUGE. The rigorous testing on a variety of datasets, revealed that ROUGE does not produce remarkable results because of its performance on both the Abstractive and Extractive methods, which is similar. Furthermore, narrowing the original reference dataset to a small field of interest, the findings are the same also considering other metrics. Moreover, a subsequent step demonstrated that multiple TS algorithm execution generally outperforms single execution. In conclusion, it is still a long way off from creating an appropriate metric to judge the summaries created by a machine.

Finally, *Chapter 5* highlights the conclusions of this research work, highlighting any future directions to follow.

Sommario della tesi

INTEGRAZIONE DEI DATI E RIEPILOGO AUTOMATICO DEL TESTO: UN PERCORSO VERSO DECISIONI COMMERCIALI PIÙ DECISIVE

Negli ultimi anni c'è stata un'esplosione dei dati condivisi on-line. La maggior parte di queste informazioni presenti su Internet è in formato testuale, e può essere utilizzata come fonte per produrre nuova conoscenza. Questi dati spesso non sono strutturati e, allo stato grezzo, non possono essere utilizzati per nessun tipo di analisi, risultando così difficili da gestire dal punto di vista dell'Information Technology (IT). Ma oltre a questi tipi di dati, la maggior parte delle aziende dispone di una vasta raccolta di dati strutturati, acquisiti e costruiti nel tempo. L'unione di queste due tipologie di informazioni rappresenta quindi una miniera d'oro per poterne trarre quanta più conoscenza possibile. Per tale motivo, il Data Pre-processing (DPP), una fase importante del processo di Data Mining, consente importanti manipolazioni sugli stessi, al fine di renderli fruibili per eventuali elaborazioni successive. I passaggi generali del DPP sono la Pulizia, l'Integrazione, la Riduzione e la Trasformazione dei dati, garantendo nel contempo la protezione della privacy. Questo lavoro di ricerca si è concentrato su due diverse applicazioni relative ai dati strutturati e non strutturati, attraverso rispettivamente un focus su Data Integration (DI) e uno sull'Automatic Text Summarization (ATS), di cui sono state esplorate le metriche di valutazione degli algoritmi. Una delle sfide più impegnative per la DI, è la ricerca di metodologie completamente o parzialmente automatiche, poiché queste tecniche spesso richiedono l'esperienza di uno specialista del dominio, in grado di dirigere il processo e migliorarne i risultati. Tuttavia, in letteratura, non sono molti gli approcci di DI completamente o parzialmente automatici, a meno che non includano esperti con specifiche competenze informatiche. Quindi, in questo studio, attraverso l'assistenza di una figura intermedia (il *Company Manager*), che non ha necessariamente competenze informatiche, utilizzando una metodologia di Information Retrieval, dei metodi di clustering e una rete neurale addestrata, abbiamo costruito un processo di DI semi-automatico. Esso è in grado sia di ridurre i conflitti persistenti nei dati, sia di garantire una visione unificata degli stessi, rispettando i vincoli originali dei dataset e fornendo un risultato di alta qualità per le valutazioni di Business Intelligence. Allo stesso tempo, avere la capacità di ridurre la quantità di testo da cui estrarre informazioni è fondamentale, quando le fonti dati coinvolte sono testuali. Ciò è importante per recuperare i concetti chiave, ma anche per velocizzare i sistemi di analisi. In particolare, l'ATS è una sfida interessante del Natural Language Processing. Il problema principale è che attualmente esistono numerosi algoritmi che provano a riassumere i documenti, utilizzando sia

tecniche statistiche (Algoritmi estrattivi) sia metodi di Artificial Intelligence (Algoritmi astrattivi). Tuttavia, per valutare la qualità dei risultati vengono utilizzate diverse metriche basate principalmente sull'analisi della sovrapposizione di *n-grammi* come la metrica ROUGE, che è quella più usata allo scopo. Pertanto, l'obiettivo del secondo argomento di ricerca è stato quello di determinare se tali metriche sono efficienti e se consentono davvero di confrontare la qualità dei risultati dei vari algoritmi di Text Summarization.

Il *Capitolo 1* presenta il contesto primario di questo lavoro di ricerca, ossia il Data Management. La panoramica iniziale prende in considerazione le varie sfide che devono essere affrontate quando si avvia un qualsiasi processo di Analisi dei dati a partire da dati strutturati e non strutturati. Vengono inoltre descritti i vari passaggi che compongono la fase di Data Pre-processing, fase essenziale per l'esecuzione corretta ed efficiente di qualsiasi processo di analisi dei dati.

Il *Capitolo 2* introduce tutte le tecniche di Machine Learning (ML) e Deep Learning (DL) utilizzate in questo lavoro di ricerca. In particolare, si focalizza sulla descrizione delle tecniche di clustering gerarchico e non gerarchico, analizzandone le differenze. Successivamente, vengono evidenziati gli approcci principali relativi alle reti neurali, facendo una panoramica dei vari tipi considerati in questo studio e, infine, vengono mostrati gli aspetti più importanti relativi all'Information Retrieval (IR), ossia il processo di ricerca delle informazioni rilevanti da una raccolta di dati.

Il *Capitolo 3* è relativo alla sfida di Data Integration. Dopo una breve presentazione dei concetti fondamentali relativi a questo particolare argomento, viene analizzato in dettaglio lo stato dell'arte attraverso una panoramica delle varie metodologie di integrazione dati. Successivamente, viene descritto il nostro nuovo approccio integrativo che unisce dati eterogenei provenienti da diverse fonti utilizzando una procedura semi-automatica. Utilizzando una tecnica di Information Retrieval, dei metodi di Clustering e una rete neurale, è possibile portare a termine il processo senza il coinvolgendo di un esperto-IT, ma solo di una figura che fungerà da collegamento tra gli sviluppatori del sistema e gli utenti finali, che non ha bisogno di avere alcuna abilità tecnica per portare a termine il lavoro. Inoltre, nel database integrato, i risultati finali manterranno tutti i vincoli tra gli attributi esistenti prima dell'integrazione.

Nel *Capitolo 4* viene approfondito l'argomento degli algoritmi di Automatic Text Summarization. Queste tecniche cercano di estrarre automaticamente informazioni importanti da uno o più testi di input, creando riassunti pur mantenendo i concetti chiave. Dopo una breve presentazione dei concetti di base correlati alla rappresentazione del testo e alla similarità fra testi, viene analizzato in dettaglio lo stato dell'arte relativo a tali algoritmi e alle metriche di valutazione degli stessi. La qualità dei riassunti prodotti da questi algoritmi è stata valutata utilizzando diverse metriche, la più popolare delle quali è ROUGE. I numerosi test effettuati su diversi dataset hanno rivelato che ROUGE

non produce buoni risultati, esaminando entrambi i metodi Astrattivi ed Estrattivi, poiché produce risultati molto simili fra loro. Inoltre, restringendo il dataset di riferimento originale a un campo di interesse ristretto, anche le ulteriori metriche analizzate risultano inefficienti. Un passaggio successivo ha dimostrato poi che la prestazione ottenuta dall'esecuzione di due algoritmi di Text Summarization in cascata, generalmente è superiore di quella ottenuta con una singola esecuzione. In conclusione, siamo ancora molto lontani dal riuscire ad ottenere una metrica adeguata per valutare i riassunti generati da una macchina.

Infine, il *Capitolo 5* mette in evidenza le conclusioni di questo lavoro di ricerca, evidenziando eventuali direzioni future da seguire.

Index