Studying, Developing, and Experimenting Contextual Advertising Systems

Alessandro Giuliani

Department of Electrical and Electronic Engineering

University of Cagliari

A thesis submitted for the degree of

PhilosophiæDoctor (PhD) degree in Electronic and Computer Engineering

March 2012

Dedicated to my parents, my sister, and my friends

Acknowledgements

I would like to thank those people that gave me the opportunity to work and enjoy during my PhD journey.

This thesis would not have been possible without the support and the insights of my advisor, prof. Giuliano Armano. I'm pleased for the given chance to reach this point of my career. I would like to thanks in particular my friend and co-tutor Eloisa Vargiu. It has been a big pleasure working with her. I have to thank her for all the support, the suggestions, the patience, and every kind of help that she gave to me, and for the person I became at this point. And of course, for the time spent out of work and her continues suggestions.

My friends and colleagues, in particular Filippo Ledda, Andrea Addis, Stefano Curatti, Nima Hatami, Ilaria Lunesu, Erika Corona, Daniele Muntoni, Emanuele Manca, Francesco Mascia, Andrea Manconi, Giuseppe De Stefanis, Zahid Akthar and Roberto Tonelli.

I would like to show my gratitude to Vanessa Murdock for her advice, collaboration, and the opportunity to work at the Yahoo! Research Labs, and Hoplo s.r.l., in particular Ferdinando Licheri and Roberto Murgia for their support.

I am also grateful to the coordinator of the PhD program prof. Alessandro Giua for his efforts, and to his collaborators Carla Piras and Maria Paola Cabasino for their helpfulness.

A special thank is dedicated to Hadi Mohammadzadeh for his support and efforts, and to everyone who has been at my side during these years.

Finally I would like to thank the only one that has been with me every day, every hour, and every moment. You have a special place in my soul. Thank you, coffee machine.

Abstract

The World Wide Web has grown so fast in the last decade and it is today a vital daily part of people. The Internet is used for many purposes by an ever growing number of users, mostly for daily activities, tasks, and services.

To face the needs of users, an efficient and effective access to information is required. To deal with this task, the adoption of Information Retrieval and Information Filtering techniques is continuously growing. Information Retrieval (IR) is the field concerned with searching for documents, information within documents, and metadata about documents, as well as searching for structured storage, relational databases, and the World Wide Web. Information Filtering deals with the problem of selecting relevant information for a given user, according to her/his preferences and interest.

Nowadays, Web advertising is one of the major sources of income for a large number of websites. Its main goal is to suggest products and services to the still ever growing population of Internet users. Web advertising is aimed at suggesting products and services to the users. A significant part of Web advertising consists of textual ads, the ubiquitous short text messages usually marked as sponsored links. There are two primary channels for distributing ads: Sponsored Search (or Paid Search Advertising) and Contextual Advertising (or Content Match). Sponsored Search advertising is the task of displaying ads on the page returned from a Web search engine following a query. Contextual Advertising (CA) displays ads within the content of a generic, third party, webpage.

In this thesis I study, develop, and evaluated novel solutions in the field of Contextual Advertising. In particular, I studied and developed novel text summarization techniques, I adopted a novel semantic approach, I studied and adopted collaborative approaches, I started a conjunct study of Contextual Advertising and Geo-Localization, and I study the task of advertising in the field of Multi-Modal Aggregation.

The thesis is organized as follows. In Chapter 1, we briefly describe the main aspects of Information Retrieval. Following, the Chapter 2 shows the problem of Contextual Advertising and describes the main contributes of the literature. Chapter 3 sketches a typical adopted approach and the evaluation metrics of a Contextual Advertising system. Chapter 4 is related to the syntactic aspects, and its focus is on text summarization. In Chapter 5 the semantic aspects are taken into account, and a novel approach based on ConceptNet is proposed. Chapter 6 proposes a novel view of CA by the adoption of a collaborative filtering approach. Chapter 7 shows a preliminary study of Geo Location, performed in collaboration with the Yahoo! Research center in Barcelona. The target is to study several techniques of suggesting localized advertising in the field of mobile applications and search engines. In Chapter 8 is shown a joint work with the RAI Centre for Research and Technological Innovation. The main goal is to study and propose a system of advertising for Multimodal Aggregation data. Chapter 9 ends this work with conclusions and future directions.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

Chapter 1

Information Retrieval

Information Retrieval (IR) might be defined as the task of finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from large collections (usually stored on computers). The term *unstructured data* refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database. In practice, almost no data are truly *structured* or *unstructured*. This is definitely true of all text data if you count the latent linguistic structure of human languages. In fact it is often desirable to facilitate IR using *semi-structured* information.

IR can cover various and heterogeneous kinds of data and information problems beyond that specified in the core definition above. More generally, an IR system does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non- existence) and whereabouts of documents relating to her/his request.

The rest of the Chapter briefly summarizes the main IR topics. Let us note that we are not interested in giving an exhaustive summary on IR, but in highlighting the main topics that could be exploited in the field of Contextual Advertising (CA). The interested reader could refer to (20) and (89) for a deep study on IR.

1.1 Indexing

According to (117), we can define an indexing operation as the task of assigning attributes to the stored data. Attributes are chosen to represent collectively the information content of the corresponding data. Given a collection D of stored data D_i, the indexing task involves the following aspects. First, it is necessary to choose a set of t

distinct attributes A_k (e.g., employee name, job classification, salary in case of personal data) to represent the information content in D. Then, for each attribute A_k, a number of different values $a_{k1}, a_{k2}, \cdots, a_{k_{n_k}}$ are defined and one of these n_k values is assigned to each record D_i for which attribute A_k applies (e.g., particular names of individual employees, particular job classifications, specific salary levels). The indexing task then generates for each stored item an index vector:

$$D_i = (a_{i1}, a_{i2}, \cdots, a_{it}) \tag{1.1}$$

where a_{ij}, also known as keyword, denotes the value of attribute A_j in item D_i. When a given a_{ij} is null, the corresponding attribute is assumed to be absent from the item description. A given attribute-value assigned to an item may be weighted or not (binary vector). In principle, a complete index vector consists of set of pairs $\langle a_{ij}, w_{ij} \rangle$:

$$D_i = (\langle a_{i1}, w_{i1} \rangle, \langle a_{i2}, w_{i2} \rangle, \cdots, \langle a_{it}, w_{it} \rangle) \tag{1.2}$$

where w_{ij} denotes the weight of keyword a_{ij}.

Given an indexed collection, it is possible to compute a similarity measure between pairs of items by comparing vector pairs. A typical measure of similarity s between D_i and D_j might be:

$$s(D_i, D_j) = \sum_{k=1}^{t} w_{ik} w_{jk} \tag{1.3}$$

For binary vectors this equals the number of matching keywords in the two vectors, whereas for weighted vectors it is the sum of the products of corresponding term weights.

As for index construction, in (135) an extensive treatment of this subject and additional indexing algorithms with different trade-offs of memory, disk space, and time, have been presented. Moffat and Bell (101) show how to construct an index *in situ*; whereas in (60) and (141) an in-depth study of index construction is discussed.

No matter what particular indexing system is used, an effective indexing vocabulary will produce a clustered object space in which classes of similar items are easily separable from the remaining items. It would be nice to relate the properties of a given indexing vocabulary directly to the clustering properties of the corresponding object space. Unfortunately, not enough is known so far about the relationship between indexing and classification to be precise on that score. The properties of normal

indexing vocabularies are related instead to concepts such as *specificity*, which denotes the level of detail at which concepts are represented in the vectors, and *exhaustivity*, which designates the completeness with which the relevant topic classes are represented in the indexing vocabulary. The implication is that specific index vocabularies lead to high precision, whereas exhaustive object descriptions lead to high recall. Attempts have been made to relate standard parameters, such as exhaustivity and specificity to quantitative measures. That formal characterization may lead to the use of optimal indexing vocabularies and the construction of optimal indexing spaces.

A widely used Text Indexing task is the bag of words representation of a text, in which each distinct term is associated with its number of occurrences in the text. The bag of words are used in most of the systems developed in this thesis.

1.2 Search and Web Search

Search is an IR task aimed at finding information stored on a computer system in response to a user's query. The search results are usually presented in a list and are commonly called hits. The most public visible form of search is Web search that searches for information on the World Wide Web.

Figure 1.1 shows a composite picture of a Web search engine composed by: (i) the crawler, which is aimed at searching for and retrieving the required information; (ii) the indexer, which is aimed at indexing the stored information; (iii) the advertisement indexer, which is aimed at indexing advertisements to be suggested to the user; and (iv) the user interface in which the user puts her/his query.

There are three broad categories into which common Web search queries can be grouped (89): (i) informational, (ii) navigational and (iii) transactional. Informational queries seek general information on a broad topic. Users with informational queries typically try to assimilate information from multiple webpages. Navigational queries seek the website or home page of a single entity that the user has in mind. In such cases, the users expectation is that the very first search result should be the home page that was in mind. Transactional query are those that are a prelude to the user performing a transaction on the Web, such as purchasing a product, downloading a file or making a reservation. In such cases, the search engine should return results listing services that provide form interfaces for such transactions.

In response to a query, a search engine can return webpages whose contents has not be (fully or even partially) indexed. Search engines generally organize their indexes in various tiers and partitions, not all of which are examined on every search. Thus,

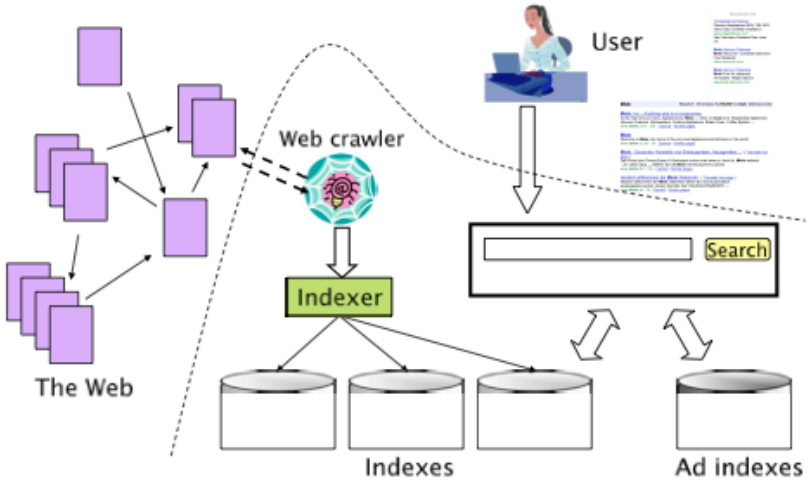

Figure 1.1: A generic architecture of a Web search engine (extracted by (89)).

search engine indexes include multiple classes of indexed pages, so that there is no single measure of index size. Thus, a number of techniques have been devised for the estimation of the ratio of the index sizes of two search engines, E_1 and E_2. The basic hypothesis underlying these techniques is that each search engine indexes a fraction of the Web chosen independently and uniformly at random. This involves two main assumptions: (i) there is a finite size for the Web from which each search engine chooses a subset and (ii) each engine selects an independent uniformly-chosen subset. Thus, an estimation technique, called the *capture recapture* method, has been defined. Suppose that we could pick a random page from the index of E_1 and test whether it is in E_2 index and symmetrically, test whether a random page from E_2 is in E_1. These experiments provide fractions x and y such that one can estimate that a fraction x of the pages in E_1 are in E_2, while a fraction y of the pages in E_2 are in E_1. Then, letting $|E_i|$ the size of the index of search engine E_i, we have $x|E_1| \approx y|E_2|$ from which we have the form:

$$\frac{|E_1|}{|E_2|} \approx \frac{x}{y} \tag{1.4}$$

If the assumption about E_1 and E_2 being independent and uniform random subsets of the Web were true, and the sampling process unbiased, then Equation 1.4 should give us an unbiased estimator for $\frac{|E_1|}{|E_2|}$. We distinguish between two scenarios: (i) either the measurement is performed by someone with access to the index of one of the search engines (e.g., an employee of E_1), or the measurement is performed by an independent

part with no access to the innards of either search engine. In the former case, a random document from one index can be simply picked. In the latter case, a random page from one search engine have to be picked from outside the search engine, thus involving to verify whether the random page is present in the other search engine.

To implement the sampling phase, a random page from the entire (idealized, finite) Web and test it for presence in each search engine might be generated. Unfortunately, picking a webpage uniformly at random is a difficult problem. Several attempts to achieve such a sample have been proposed, let us recall here:

- *Random searches.* It starts with a search log of Web searches and sends a random search from this log to E_1 and a random page from the results. Since such logs are not widely available outside a search engine, one implementation is to trap all search queries going out of a work group that agrees to have all its searches logged.

- *Random IP addresses.* This technique generates random IP addresses and send a request to a Web server residing at the random address, collecting all pages at that server.

- *Random walks.* It runs a random walk starting at an arbitrary webpage. This walk would converge to a steady state distribution, from which we could in principle pick a webpage with a fixed probability.

- *Random queries.* It picks a page (almost) uniformly at random from a search engine's index by posing a random query to it. This approach is noteworthy for two reasons: it has been successfully built upon for a series of increasingly refined estimates, and conversely it has turned out to be the approach most likely to be misinterpreted and carelessly implemented, leading to misleading measurements.

The estimation of Web search index sizes has a long history of development covered, for instance, in (26) (77) (61) (115) (21).

Search is also applied to CA, in which the advertising approach is viewed as a problem of query expansion and rewriting, as proposed by Murdock et al. (102) and recalled in Chapter 2. A preliminary study related to mobile query logs for a geo-targeted advertising is performed in Chapter 7

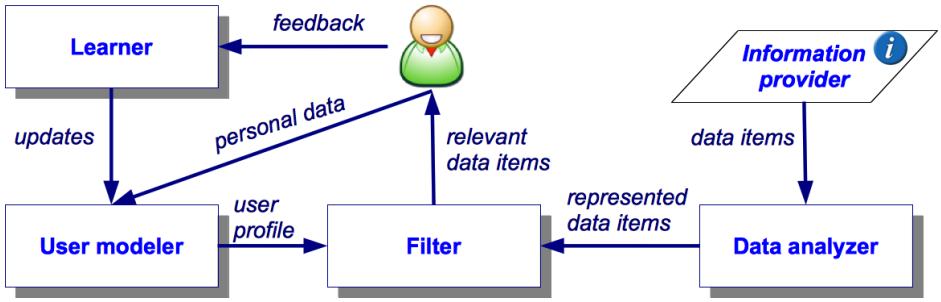

Figure 1.2: A generic model of IF systems (re-drawn from (56)).

1.3 Information Filtering

Information filtering is aimed at presenting only relevant information to users. Information filtering systems cover a broad range of domains, technologies, and methods involved in the task of providing users with the information they require (as described by Hanani et al. (56)). Information filtering systems are, typically, able to handle large amounts of data with the following characteristics: (i) primarily textual, (ii) unstructured or semi-structured, and (iii) based on user profiles. Information filtering systems are also expected to remove irrelevant data from incoming streams of data items.

According to (56), a generic information filtering system includes four basic components: data analyzer, user modeler, filter, and learner (see Figure 1.2). The data analyzer obtains or collects data items from information providers. Data items are analyzed and represented in a suitable format (e.g., as a vector of indexed terms). The user modeler, which explicitly and/or implicitly gathers information about users and their information needs, constructs user models (e.g., user profiles). The filter matches the user model with the represented data items and decides if a data item is relevant to the user. Sometimes the decision is binary (i.e., relevant or irrelevant) and sometimes is probabilistic (i.e., the data item gets a relevance rank). The filtering process can be applied to a single data item (e.g., a new blog post) or to a set of data items (e.g., a website). The user who gets the suggested data item is the only one that can state the real relevancy. Her/his evaluation enables further feedback, to be provided to the learner. The learner is aimed at improving the filtering ability, with the goal of dealing with the difficulties that arise from modeling users and from changes in their information needs. In particular, the learning process is able to counteract shifts in users' interests by updating their model to avoid inaccuracies that occur in profiles affecting filtering results.

In IR, information filtering is typically performed by Recommender Systems (RS).

They have been widely advocated as a way of coping with the problem of information overload in many application domains and, to date, many recommendation methods have been developed (133). The recommendation problem can be formulated as follows: let U be the set of all users and let I be the set of all possible items that can be recommended (e.g., books, movies, and publications). Let f be a utility function that measures the usefulness of item i to user u, i.e. $f : U \times I \rightarrow R$, where R is a totally ordered set (e.g., non-negative integers or real numbers within a certain range). Then, for each user $u \in U$ we want to choose such item $i' \in I$ that maximizes the user's utility. More formally:

$$\forall u \in U : i'_u = \underset{i \in I}{argmax} \;\; f(u, i) \tag{1.5}$$

A RS could be used to perform a CA task, and vice versa. In chapter 6.1 a unifying view of Recommendation and CA is proposed.

1.4 Text Categorization and Hierarchical Text Categorization

Text categorization (also known as text/document classification) is the task of assigning predefined categories to text documents. It can provide conceptual views of document collections and has many important applications in the real world. Many document collections are useful to be categorized into classes, for instance: news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects; in spam filtering, email messages are classified into the two categories of spam and non-spam, respectively.

More formally, text categorization is the task of assigning a boolean value to each pair $\langle d_j, c_i \rangle \in D \times C$, where D is a domain of documents and $C = \{c_k \mid k = 1, 2, ..., N\}$ is a set of N predefined categories. A value of T assigned to $\langle d_j, c_i \rangle$ indicates a decision to file d_j under c_i, while a value of F indicates a decision not to file d_j under c_i. Thus, the task is to approximate the unknown function $\hat{\Phi} : D \times C \rightarrow \{T, F\}$ (that describes how documents ought to be classified) by means of a function $\Phi : D \times C \rightarrow \{T, F\}$ called the classifier (also known as rule, or hypothesis, or model) such that $\hat{\Phi}$ and Φ *coincide as much as possible* (120).

Text categorization appeared as a research area in the 1960s (91). Only in the 1990s it became a major field in information science due to the increased interest in

its diverse applications such as document indexing with controlled vocabulary, filtering of irrelevant information, webpage categorization, email management, detection of text genre, and many others. Until the mid-1990s researchers mostly ignored the hierarchical structure of categories that occur in several domains. In 1997, Koller and Sahami (72) carried out the first proper study on hierarchical text categorization on the Reuters-22173 collection. Documents were classified according to the given hierarchy by filtering them through the single best-matching first-level class and then sending them to the appropriate second level. According to (67), hierarchical text categorization is a text categorization task performed relying on a given taxonomy $TAX = \langle C, \leq \rangle$, where $C = \{c_k \mid k = 1, 2, ..., N\}$ is a set of N predefined categories and "$\leq$" is a reflexive, anti-symmetric, and transitive binary relation. In the most general case, TAX can be though of as a strict partially ordered set (strict poset), which can be graphically represented by a directed acyclic graph (DAG).

In this work a Text Classification task is performed to extract semantic information form a textual document. The task is performed by the adoption of classifier based on the Rocchio approach, as discussed in Chapter 5.

1.5 Text Mining

Text Mining is an IR task aimed at discovering new, previously unknown information, by automatically extracting it from different text resources (23). The term "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information (120).

Automatic extraction of metadata (e.g., subjects, language, author, key-phrases) is a prime application of text mining techniques. Although, contemporary automatic document retrieval techniques bypass the metadata creation stage and work on the full text of the documents directly, text mining has been largely applied to learn metadata from documents. Language identification is a relatively simple mining task aimed at providing an important piece of metadata for documents in international collections. A simple representation for document categorization is to characterize each document by a profile that consists of the "n-grams", or sequences of n consecutive letters, that appear in it. Documents are preprocessed by splitting them into word tokens containing letters and apostrophes, padding each token with spaces, and generating all possible n-grams of length 1 to 5 for each word in the document. These n-grams are counted and sorted into frequency order to yield the document profile. An alternative approach is

to use words instead of n-grams and compare occurrence probabilities of the common words in the language samples with the most frequent words of the test data. Author metadata is one of the primary attributes of most documents. It is usually known and need not be mined, but in some cases authorship is uncertain and must be guessed from the document text. Authorship ascription is often treated as a text mining problem. In the scientific and technical literature, keywords and key-phrases are attached to documents to give a brief indication of what they are about. Key-phrases are a useful form of metadata because they condense documents into a few pithy phrases that can be interpreted individually and independently of each other. Given a large set of training documents with key-phrases assigned to each, text mining techniques can be applied to assign appropriate key-phrases to new documents.

In contrast to text categorization, document clustering is an "unsupervised" learning technique in which there is no predefined category or class, but groups of documents that belong together are sought. For example, document clustering assists in retrieval by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query (92). Although they do not require training data to be pre-classified, clustering techniques are generally far more computation-intensive than supervised schemes (134). Nevertheless, clustering has been largely applied in text mining applications. Trials of unsupervised schemes include the work by: Aone et al. (11) who use the conceptual clustering scheme COBWEB (50) to induce natural groupings of close-captioned text associated with video news feeds; Liere and Tadepalli (81) who explore the effectiveness of AutoClass (37) in producing a classification model for a portion of the Reuters corpus; and Green and Edwards (54) who use AutoClass to cluster news items gathered from several sources into *stories*, which are groupings of documents covering similar topics.

One of the main subfields of text mining is information extraction, i.e., the task of filling templates from natural language input (12). Typical extraction problems address simple relationships among entities, such as finding the predicate structure of a small set of pre-determined propositions. Machine learning has been applied to the information extraction task by seeking pattern-match rules that extract fillers for slots in the template (126) (63) (33) (52). The extracted information can be used in a subsequent step to learn rules that characterize the content of the text itself.

Text mining is also applied to provide summaries of documents or group of documents. Several studies and novel applications will be discussed in Chapter 4.

Chapter 2

Contextual Advertising

The main research field of this thesis is Contextual Advertising, a form of promotion that uses the World Wide Web with the purpose of suggesting marketing messages (usually textual ads) in a webpage, with the goal of capture the users' interests in order to interact with the ads and generate revenue. In this chapter we introduce the background and the state-of-the-art of CA.

2.1 Online Advertising

Online Advertising is an emerging research field, at the intersection of IR, Machine Learning, Optimization, and Microeconomics. Its main goal is to choose the right ads to present to a user engaged in a given task.

Online advertising has come a long way in the past decade. When the Internet was first breaking out into the mainstream, and becoming accessible to regular consumers and not just academics and scientists, there was still a lot of doubt about how viable it could be as a commercial medium. Many early attempts at online advertising met with limited success, at best. This was also compounded with the general problem that many advertisers simply did not understand the online space, and still looked at advertising from the perspective of the print or television mediums.

The first approaches in Online Advertising were simple in terms of the way there were presented to a generic user. The common task of advertising were simply based on a payment by the advertiser to have a small banner ad placed on a website for a given period of time. In this rudimentary way the advertisers had to depend on the website's statistics to get an idea of how many people actually saw their ad. In this way Online Advertising could be seen as a "print advertising", based solely on the dimension of

ads, and the prices were related on the expected number of unique website visitors who would see the ad.

The earliest advanced systems were developed in the mid-90's, and they used to allow multiple ad banners to rotate in the same section of the webpage, and kept track of the amount of impressions for each ad. Some ad services made available the tracking of clickthroughs, so advertisers could find out the percentage of users really interested in the proposed ad. These first innovations made it possible to sell online advertising in new ways. An advertiser could specify that they only wanted to buy 100,000 impressions in a month, for example. With the capability to rotate banners, publishers also could be able to sell ad space to multiple advertisers, rather than offer to just one advertiser at a time. Online advertising became to be different respect to the ad services offered by print and television sources.

Online Advertising had a fast growth in development of systems and tools increasingly efficient and effective, with the adoption of more and more suitable methods and approaches. Nowadays Online Advertising supports a large swath of today's Internet ecosystem.

There are two primary channels for distributing ads: Sponsored Search (or Paid Search Advertising) and CA (or Content Match). Sponsored Search advertising is the task of displaying ads on the page returned from a Web search engine following a query. CA displays ads within the content of a generic, third party, webpage. Sponsored Search (SS), also called paid search advertising, displays ads on the page returned from a Web search engine following a query. CA displays ads within the content of a generic, third party, webpage. A commercial intermediary, namely ad network, is usually in charge of optimizing the selection of ads with the twofold goal of increasing revenue (shared between publisher and ad network) and improving user experience. In other words, CA is a form of targeted advertising for ads appearing on Web sites or other media, such as content displayed in mobile browsers. The ads themselves are selected and served by automated systems based on the content displayed to the user.

The SS market developed quicker than the CA market. In SS a generic ad is usually represented by its "bid phrases" that are the queries for which the advertiser would like to display its ad (49). The evolution of advertising lead to a significant diffusion of the CA services, that are prevalent in several fields. As a matter of fact, today most of the for-profit non-transactional Web sites (that is, sites that do not sell anything directly) rely on revenue of CA. The prevalent pricing model for textual ads is the pay-per-click (PPC), in which a certain amount is paid by the advertiser for each user's click on his advertisement. There are also other models, such as, pay-per-impression

(PPI), where the advertiser pays for the number of exposures of an ad, and pay-per action where the advertiser pays only if the ad leads to a revenue. Many techniques used in CA to place ads on webpages may be used to place ads in response to a user's query, as in SS. SS can be thought as a document retrieval problem, where ads are documents to be retrieved in response to a query. Ads can be partially represented by a set of relevant keywords. Carrasco et al. (34) approached the problem of keyword suggestion by clustering bipartite advertiser-keyword graphs. Joachims (65) proposed to use click-data to learn ranking functions for results of a search engine as an indicator of relevance. Ciaramita et al. (39) studied an approach to learn and evaluate Sponsored Search systems based solely on click-data, focusing on the relevance of textual content.

2.2 Background

CA is a form of targeted advertising for ads appearing on websites or other media, such as content displayed in mobile browsers. As discussed in the work of Broder et al. (30), CA is an interplay of four players:

- The *advertiser* provides the supply of ads. As in traditional advertising, the goal of the advertisers can be broadly defined as the promotion of products or services.
- The *publisher* is the owner of the webpage (*target page*) on which the advertising is displayed. The publisher typically aims to maximize advertising revenue while providing a good user experience.
- The *ad network* is a mediator between the advertiser and the publisher; it selects the ads to display on the webpages. The ad network shares the advertisement revenue with the publisher.
- The *Users* visit the webpages of the publisher and interact with the ads, by clicking and visiting the ad page (*landing page*).

Figure 2.1 sketches the interactions among the four players in a typical CA task.

The ads themselves are selected and served by automated systems based on the content displayed to the user. In fact, given a webpage, it is more suitable to place ads related to the content of the webpage, in order to improve the user's click probability. Considering a generic user that visit a webpage with a specific topic, it is logical to think that the user should be more interested in services topically related to the visited page. For instance, if the webpage in which the ad should be placed is concerned with fashion content, it is intuitively more appropriate provide ads related to fashion services

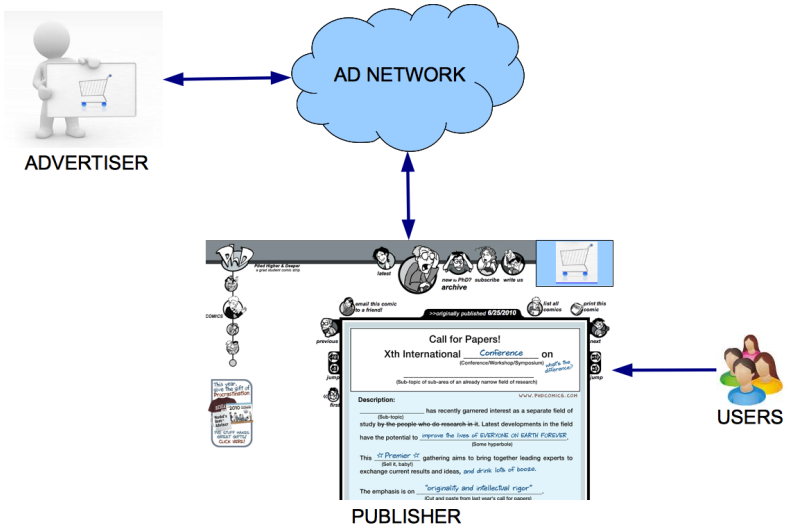

Figure 2.1: The four players in a CA task.

or products, rather than sports or traveling services. Several studies have proven that topical advertising increase the number of ad-clicks (36).

2.2.1 The Problem

To formulate the CA problem, let P be the set of webpages and let A be the set of ads that can be displayed. The revenue of the network, given a page p, can be estimated as:

$$R = \sum_{i=1}^{k} Pr(click|p, a_i) \cdot price(a_i, i) \tag{2.1}$$

where k is the number of ads displayed on page p and $price(a_i, i)$ is the click-price of the current ad a_i at position i. The price in this model depends on the set of ads presented on the page. Several models have been proposed to determine the price, most based on generalizations of second price auctions. For the sake of simplicity, we ignore the pricing model and reformulate the CA problem as follows: for each page $p \in P$ we want to select the ad $a' \in A$ that maximizes the click probability. Formally:

$$\forall p \in P : a'_p = \underset{a \in A}{argmax} \;\; Pr(click|p, a) \tag{2.2}$$

An alternative way to formulate the CA problem is the following: let P be the set of webpages and let A be the set of ads that can be displayed. Let f be a utility function that measures the matching of a to p, i.e. $f : P \times A \rightarrow R$, where R is a totally ordered set (e.g., non-negative integers or real numbers within a certain range). Then, for each page $p \in P$ we want to select $a' \in A$ that maximizes the page utility function. More formally:

$$\forall p \in P : a'_p = \underset{a \in A}{argmax} \;\; f(p, a) \tag{2.3}$$

Note that in this case the utility function f can also be viewed as an estimation of the probability that the corresponding ad be clicked.

2.2.2 Related Work

A natural extension of search advertising consists of extracting phrases from the target page and matching them with the bid phrases of ads. Yih et al. (139) proposed a system for phrase extraction, which uses a variety of features to determine the importance of page phrases for advertising purposes. To this end, the authors proposed a supervised approach that relies on a training set built using a corpus of pages in which relevant phrases have been annotated by hand.

Ribeiro-Neto et al. (111) examined a number of strategies to match pages and ads based on extracted keywords. They represented both pages and ads in a vector space and proposed several strategies to improve the matching process. In particular, the authors explored the use of different sections of ads as a basis for the vector, mapping both page and ads in the same space. Since there is a discrepancy between the vocabulary used in the pages and in the ads (the so called *impedance mismatch*), the authors improved the matching precision by expanding the page vocabulary with terms from similar pages.

In a subsequent work, Lacerda et al. (75) proposed a method to learn the impact of individual features by using genetic programming. The results showed that genetic programming helps to find improved matching functions.

Broder et al. classified both pages and ads according to a given taxonomy and matched ads to the page falling into the same node of the taxonomy. Each node is built as a set of bid phrases or queries corresponding to a certain topic. Results showed a better accuracy than that corresponding to the classic systems (i.e., systems based on syntactic matching only). Let us also note that, to improve performances, this system may be used in conjunction with more general approaches.

Nowadays, ad networks need to deal in real time with a large amount of data, involving billions of pages and ads. Hence, efficiency and computational costs are crucial factors in the choice of methods and algorithms. Anagnostopoulos et al. (9) presented a methodology for Web advertising in real time, focusing on the contributions of the different fragments of a webpage. This methodology allows to identify short but informative excerpts of the webpage by means of several text summarization techniques, used in conjunction with the model developed by Broder.

Since bid phrases are basically search queries, another relevant approach is to view CA as a problem of query expansion and rewriting. Murdock et al. considered a statistical machine translation model to overcome the problem of the impedance mismatch between pages and ads. To this end, they proposed and developed a system able to re-rank the ad candidates based on a noisy-channel model. In a subsequent work, Ciaramita et al. (38) used a machine learning approach, based on the model described in the work of Broder, to define an innovative set of features able to extract the semantic correlations between the page and the ad vocabularies.

2.3 State of the art

The evolution of the ad network has lead to effective and efficient approaches that have been the basis for the current ad networks. In this section two fundamental aspects of modern CA are shown.

2.3.1 Syntactics and semantics

Several studies estimated the ad relevance between a webpage and the ad by analyzing the co-occurrence of the same terms or phrases within the page and within the ad (as proposed by Lacerda and Ribeiro-Neto). However the approaches based solely on syntactic analyses of the original text can lead to poor performances of ad networks. Polysemy is one of the major culprits for many irrelevant selected ads; for example the term "bass" could referred to a kind of fish, to a musical instrument, or to an audio tone. Furthermore, there are several reasons for a non-performative syntactic ad network, mostly due to the lack of a content analysis. A famous example is about a suggested ad in a page related to a news item about a headless body found in a suitcase; the proposed ad was a luggage of a well known brand.

In order to solve these problems, the state-of-the-art ad networks rely on approaches that combine syntactic and semantic analyses. The semantic analysis is based on the adoption of external lexical resources (typically, a taxonomy of topics), which is devoted

to classify pages and ads to extract suitable features to be used in conjunction with the classic syntactic keywords. Furthermore, the adoption of a hierarchical taxonomy allows for a gradual generalization of the ad space, in case of no relevant ads are found for the precise page topic. For example, let us consider a page related to the class "snowboarding", and no ads were found in the repository; the system would rank ads belonging to the class of "skiing" as high relevant ads. The class "skiing" is in fact a sibling of "snowboarding", and both have the parent node "winter sport". The example is appropriate if we think about an ad offering winter dresses. In some sense, the taxonomy topics are used to extract the set of candidate ads and the keywords to narrow down the search to items with a smaller granularity. Usually, the taxonomy contains topics that do not change fast (e.g., a brand of mobile phones), whereas the keywords capture more specific topics (e.g., a specific model of mobile phone for the selected brand) that cannot be included in the taxonomy, because updating it each time a new product (e.g., a new model of mobile phone) becomes available in the market is prohibitively expensive when dealing with millions of products.

2.3.2 Real Time Advertising

The classic approaches of CA relied on a prior analysis of the entire content of the webpage, due to the fact that the webpages were mostly static, but nowadays these approaches are mostly useless. This is due to the evolution of Web, which has lead to a high level of dynamicity for several factors. Let us consider the actual trend in the Web, where daily a continuous growing amount of people use that several times for many purposes, for useful activities or just for fun. A generic webpage could be updated several times in a short span of time (for example, a blog or a personal page of a social network). Furthermore, the current approaches for build a webpage rely on dynamic tools and scripts, for which the content of the page is dynamically created at the same time of the user's browser request. Another aspect is that many pages belong to the *invisible web*, that is, they are inaccessible with the common Web-crawlers or spiders used by the search engines. Moreover, some pages require authorizations or cookies that are located in users' computers, not in the servers. These crucial factors preclude that a page could be processed ahead of time, and the ads needs to be matched with the page while it is being served to the user. This real time process limits the amount of time allotted for the analysis.

To reduce the computational times, the analysis of the entire body of the page lead to the need of high amount resources, in terms of computational and communication tools and infrastructures, that entail prohibitive communication, computational, and

latency costs. One of the most suitable method to reduce these costs, commonly used in the CA field, is a Text Summarization process, in order to craft in real time short page summaries from the webpage. Empirical evaluation proves that the matching between the ads and the summaries does not significantly decreases the relevance of the suggested ads, but the performances are comparable with the matching between ads and the entire textual body of the page. Anagnostopoulos et al. has proven that considering only a carefully selected 5% fraction of the page sacrifices at most 3% in ad relevance.

Chapter 3

Techniques and Datasets to Perform Experiments on Contextual Advertising Systems

The research activity relied on some common tools, infrastructures, algorithms, and data used in the proposed systems and applications. This Chapter is devoted to the description of:

- the baseline system used to compare the performances of the studied, developed, and implemented systems;
- the evaluation metrics used to evaluate the performances;
- the adopted datasets.

3.1 The Baseline System

For the experiments of most of the developed systems and algorithms we proposed and implemented a CA system compliant with the state-of-the-art.

Figure 3.1 sketches the model which illustrates a generic architecture that can give rise to specific systems depending on the choices made on each involved module.

The three modules are roughed out as follows:

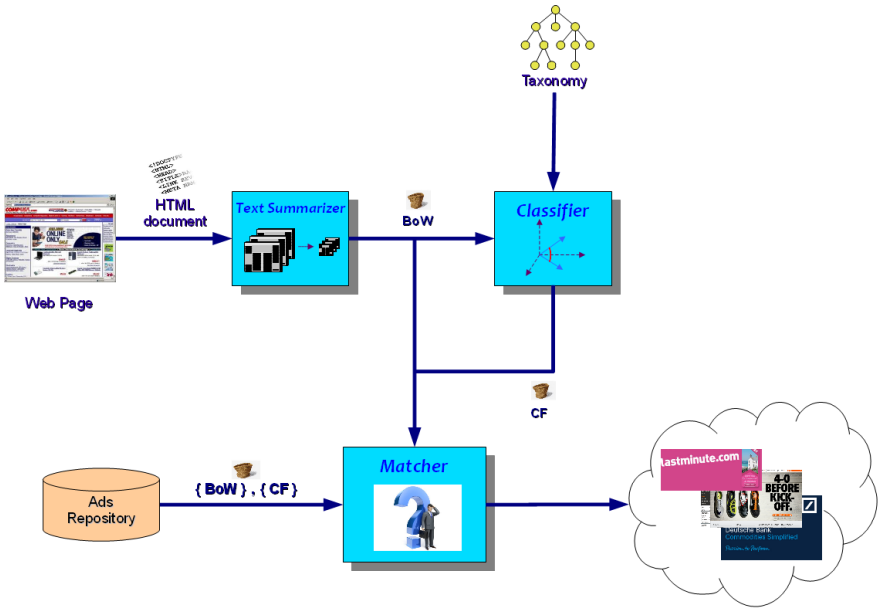

Figure 3.1: The Baseline System.

Text summarizer

Its main purpose is to transform an HTML document (a webpage or an ad) into an easy-to-process document in plain-text format, while maintaining important information. The text summarizer selects meaningful short excerpts of the textual content to reduce the data flow. It outputs a vector representation of the original HTML document as bag of words (BoW), filtered by removing non meaningful terms and by stemming the selected terms. The output is represented by adopting the vector space model, in which each document is a vector with one real-valued component for each term. Each term is weighted by TF-IDF (118). The coded page is processed by the sub modules sketched in Figure 3.2

- ***Parser.*** Given the coded page, this module is devoted to the extraction of textual content, obtained by removing tags, comments, meta-data, and every non textual item[1]. The module outputs the full content text of the webpage.

- ***Tokenizer.*** It is devoted to tokenize the text, and each token is a paragraph of the input text (the title of the page is considered as a token itself).

[1]To this end, the Jericho API for Java has been adopted, described at the webpage: http://jericho.htmlparser.net/docs/index.html

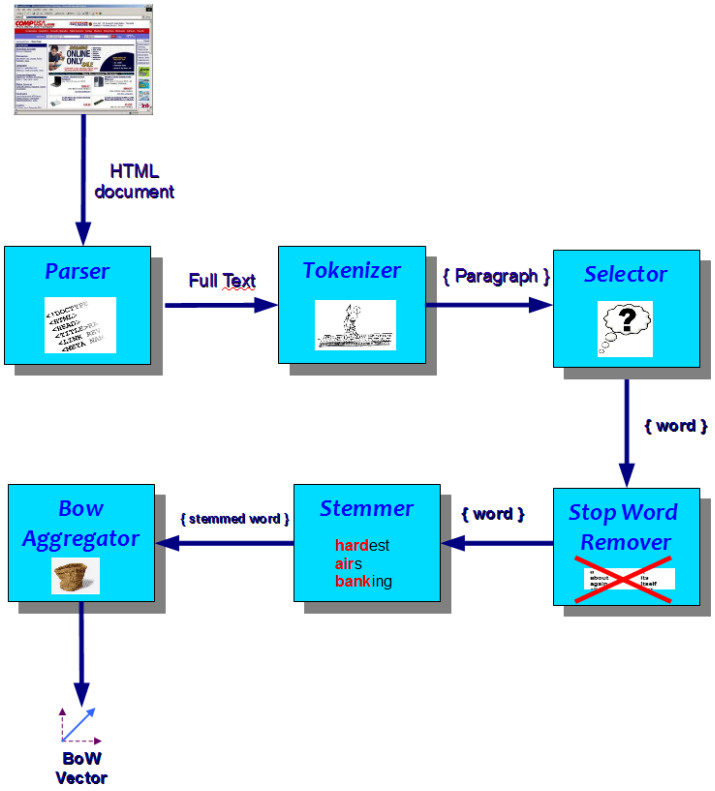

Figure 3.2: Text Summarizer.

- ***Selector.*** It is aimed to select the paragraphs as a summary, depending on the adopted technique as described in Chapter 4.

- ***Stop Word Remover.*** It performs the task of removing non meaningful terms from text (such as articles or conjunctions) to obtain a list of only useful words.

- ***Stemmer.*** It performs the process of stemming, in order to extract the root for each term. The adopted technique is the well know Porter's algorithm.

- ***BoW aggregator.*** It gives a vector-based representation of the stemmed words, first by aggregating the bag of words of them, and the projecting the bag of words in the vector space.

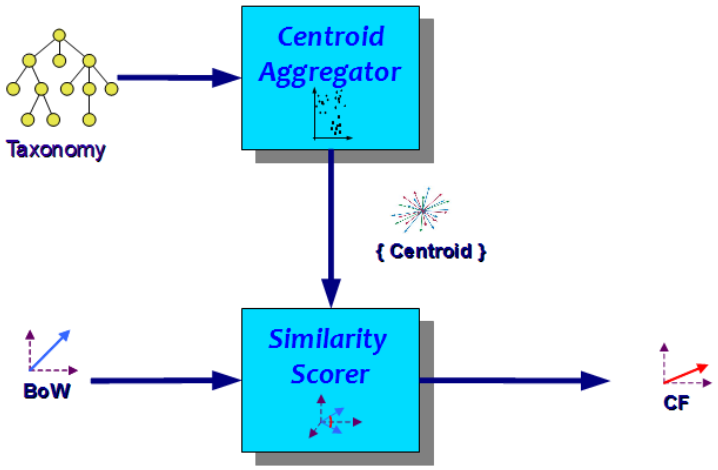

Figure 3.3: Classifier.

Classifier

Text summarization is a purely syntactic analysis and the corresponding Web-page classification is usually inaccurate. To alleviate possible harmful effects of summarization, both page excerpts and advertisings are classified according to a given set of categories. The corresponding classification-based features (CF) are then used in conjunction with the original BoW. In the current implementation, we adopt a centroid-based classification technique (55), which represents each class with its centroid calculated starting from the training set. A page is classified measuring the distance between its vector and the centroid vector of each class by adopting the cosine similarity.

Figure 3.3 shows the main modules of the classifier.

- ***Centroid Aggregator*** It is devoted to the training phase of the classifier. For each node of the input taxonomy it calculates the centroid starting from the training pages, represented in the vector space. The training pages are preprocessed as a generic test page by the text summarizer. The module outputs the list of centroid vectors.

- ***Similarity Scorer*** It takes in input the page to be analyzed, in the BoW vector representation, and for each centroid vector it calculates the similarity between

the page and the centroid. The output is a vector which its features are the similarity scores for each centroid. The output vector is the classification-based features.

Matcher

It is devoted to suggest ads (a) to the webpage (p) according to a similarity score based on both BoW and CF. In formula (α is a global parameter that permits to control the emphasis of the syntactic component with respect to the semantic one):

$$score(p, a) = \alpha \cdot sim_{BoW}(p, a) + (1 - \alpha) \cdot sim_{CF}(p, a) \tag{3.1}$$

where, $sim_{BoW}(p, a)$ and $sim_{CF}(p, a)$ are cosine similarity scores between p and a using BoW and CF, respectively.

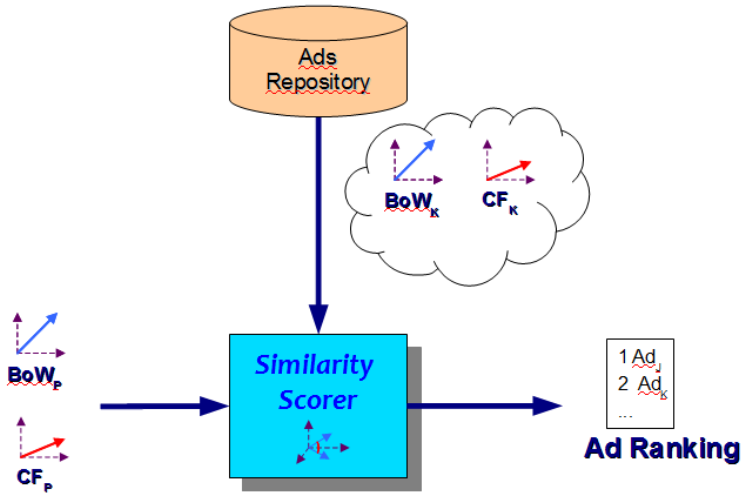

Figure 3.4: Matcher.

Figure 3.4 shows the matcher module. It takes in input both bag-of-words and classification features vectors of the test page (BoW_P and CF_P respectively) and the set of ads (BoW_K and CF_K for the $k-th$ ad). For each ad, it calculates the score with the formula 3.1. The output is the ad ranking list, and the system selects the first ads of the list, depending on how many messages should be placed in the target place.

3.2 The adopted Datasets

Experiments have been performed on two datasets extracted by the Open Directory Project and Yahoo! Categories.

BankSearch dataset

The BankSearch Dataset (124) is built using the Open Directory Project and Yahoo! Categories[1], consisting of about 11000 webpages classified by hand in 11 different classes.

Figure 3.5 shows the overall hierarchy. The 11 selected classes are the leaves of the taxonomy, together with the class *Sport*, which contains Web documents from all the sites that were classified as *Sport*, except for the sites that were classified as *Soccer* or *Motor Sport*. The authors show that this structure provides a good test not only for generic classification/clustering methods, but also for hierarchical techniques.

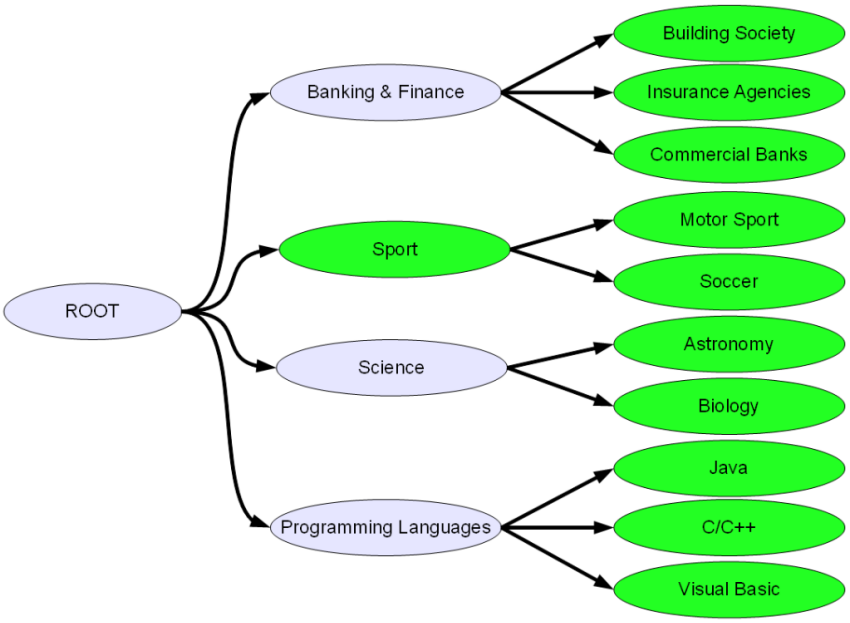

Figure 3.5: Class hierarchy of BankSearch Dataset.

[1]http://www.dmoz.org and http://www.yahoo.com, respectively

Recreation dataset

It is a self made dataset, and it consists of about 5000 webpages classified by hand in 18 categories (see Figure 3.6). As for the Banksearch dataset, the selected pages are distinct for each category.

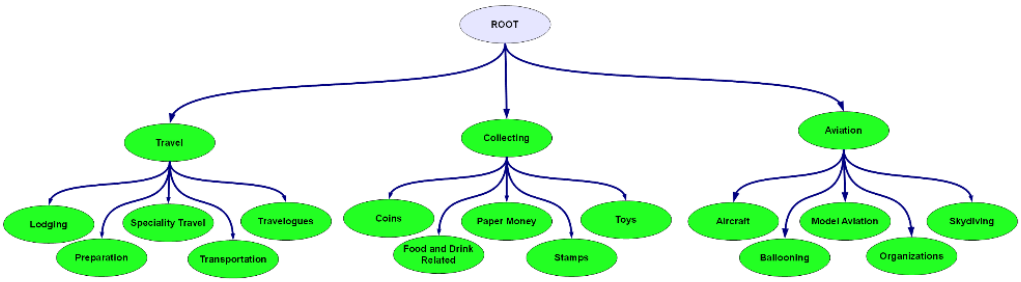

Figure 3.6: The taxonomy of Recreation Dataset.

3.3 Evaluation metrics

Relevance Score

In CA, the relevance of the suggested ad for a given page is evaluated by human assessors, that usually provide a judgment based on three score levels: *relevant, somewhat relevant, and irrelevant*. According to this task, we propose an automatic evaluation algorithm, based on the associated categories for both page and ad. The categories should belong to a given taxonomy.

Given a page p and an ad a, the $\langle p,\ a\rangle$ pair has been scored on a 1 to 3 scale defined as follows (see Figure 3.7):

1 - Relevant. a is semantically directly related to the main subject of p. For example, if p is about *how to build the perfect doll house* and a is a doll house shop, the $\langle p,\ a\rangle$ pair will be scored as 1. In other words, according to Figure 3.7-a, both p and a belong to the same class C_5 (*Toys*).

2 - Somewhat relevant. Three cases may occur: (i) a is related to a similar subject of p (*sibling*); (ii) a is related to the main topic of p in a more general way (*generalization*); or (iii) a is related to the main topic of p in a too specific way (*specialization*). As an example of *sibling*, let us assume that p is about *how to*

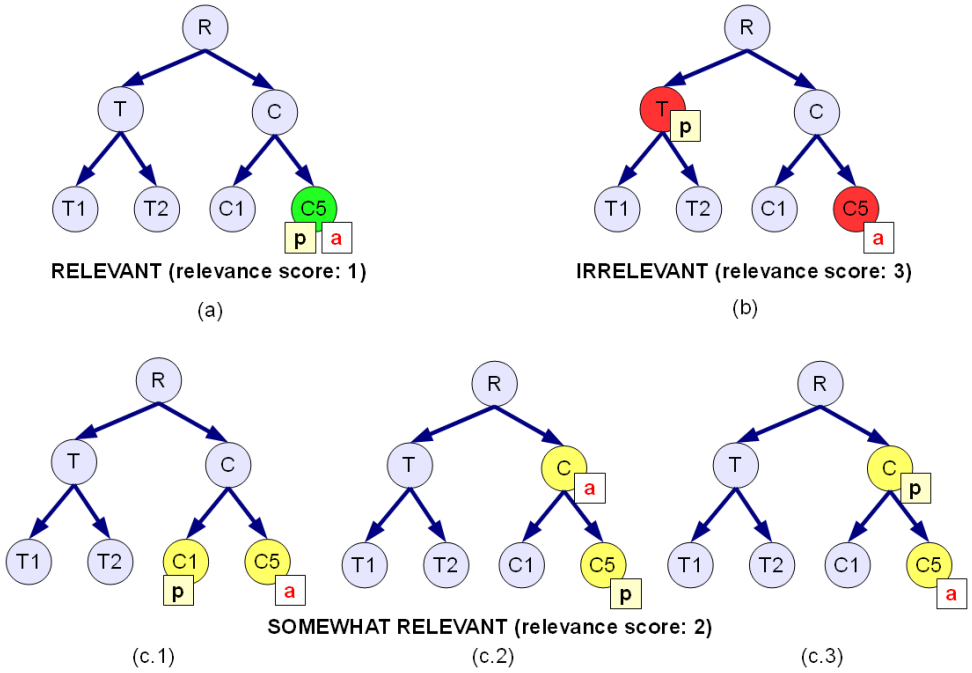

Figure 3.7: The adopted relevance scores.

build the perfect doll house and a is about collecting coins, the $\langle p,\ a\rangle$ pair will be scored as 2. In other words, according to Figure 3.7-c.1, p belongs to the class C_5 (*Toys*) and a belongs to its sibling class C_1 (*Coins*). As an example of *generalization*, let us assume that p is about *how to build the perfect doll house* and a is a shop that sells several collecting items, the $\langle p,\ a\rangle$ pair will be scored as 2. In other words, according to Figure 3.7-c.2, p belongs to the class C_5 (*Toys*) and a to its super-class C (*Collecting*). Finally, as an example of *specialization*, let us assume that p is about *collecting items from all over the world* and a is a shop that sells doll houses, the $\langle p,\ a\rangle$ pair will be scored as 2. In other words, according to Figure 3.7-c.3, p belongs to the class C (*Collecting*) and a to its sub-class C_5 (*Toys*).

3 - Irrelevant. The ad is unrelated to the page. For example, if p is about *how to build the perfect doll house* and a is a travel agency, the $\langle p,\ a\rangle$ pair will be scored as 3. In other words, according to Figure 3.7-b, p belongs to the class C_5 (*Toys*) and a to the class T (*Travel*), i.e., to a different branch of the taxonomy.

According to state-of-the-art CA systems, we considered as True Positives (TP) pairs scored as 1 or 2, and a False Positives (FP) pairs scored as 3, as schematized in

the Figure 3.8.

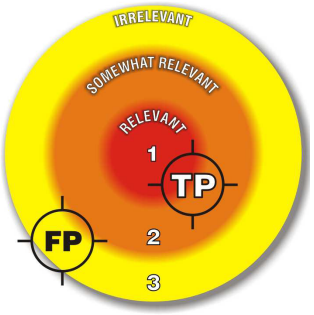

Figure 3.8: Schema for true positives and false positives.

Precision

We calculate the precision π of the considered systems in a classical way:

$$\pi = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FP_i)} \tag{3.2}$$

where N is the total number of pages. In particular, five different experiments have been performed for each system, the number of suggested ads ranging from 1 to 5. Results are then calculated in terms of the precision $\pi@k$, with $k \in [1, 5]$:.

$$\pi@k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{k} TP_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{k} (TP_{ij} + FP_{ij})} \tag{3.3}$$

3. TECHNIQUES AND DATASETS TO PERFORM EXPERIMENTS ON CONTEXTUAL ADVERTISING SYSTEMS

Chapter 4

The Impact of Text Summarization on Contextual Advertising

The first step of the research activity has been the study of the impact of the syntactic phase on CA. To this end, we considered the Text Summarizer module shown in chapter 2, and focused on the Text Summarization techniques, studying the classic techniques and proposing novel methods, with comparative experiments in order to evaluate the effectiveness of our proposals. Then, we evaluated the impact of the Text Summarization on CA, evaluating a system implemented for each technique.

G.Armano, **A.Giuliani**, E.Vargiu. *Experimenting Text Summarization Techniques for Contextual Advertising*, IIR'11: Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop, 2011.

G.Armano, **A.Giuliani**, E.Vargiu. *Studying the Impact of Text Summarization on Contextual Advertising*, TIR'11: Proceedings of the 8th International Workshop on Text-based Information Retrieval, 2011.

G.Armano, **A.Giuliani**, E.Vargiu. *Using Snippets in Text Summarization: a Comparative Study and an Application*, IIR'12: 3rd Italian Information Retrieval (IIR) Workshop, 2012.

4.1 Background

During the '60, a large amount of scientific papers and books have been digitally stored and made searchable. Due to the limitation of storage capacity, documents were stored, indexed, and made searchable only through their summaries (42). So that, how to automatically create summaries became a primary task and several techniques were defined and developed (48; 85; 116).

Radev et al. (107) define a summary as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". This simple definition highlights three important aspects that characterize research on automatic summarization: (i) summaries may be produced from a single document or multiple documents; (ii) summaries should preserve important information; and (iii) summaries should be short. Unfortunately, attempts to provide a more elaborate definition for this task resulted in disagreement within the community (41).

More recently, there was a renewed interest on automatic summarization techniques. The problem now is no longer due to limited storage capacity, but to retrieval and filtering needs. In fact, since digitally stored information is more and more available, users need suitable tools able to select, filter, and extract only relevant information. Therefore, text summarization techniques are currently adopted in several fields of IR and filtering (20), such as, information extraction (109), text mining (135), document classification (123), RS (112), and CA.

Automatic text summarization is a technique in which a text is summarized by a computer program. Given a text, its summary, which is a non redundant extract from the original text, is returned.

Simple summarization-like techniques have been long applied to enrich the set of features used in text categorization. For example, a common strategy is to give extra weight to words appearing in the title of a story (100) or to treat the title-words as separate features, even if the same words were present elsewhere in the text body (47). It has been also noticed that many documents contain useful formatting information, loosely defined as context, that can be utilized when selecting the salient words, phrases or sentences. For example, Web search engines select terms differently according to their HTML markup (22). Summaries, rather than full documents, have been successfully applied to document clustering (53). Ker and Chen (66) evaluated the performance of a categorization system using title-based summaries as document descriptors. In their experiments with a probabilistic TF-IDF based classifier, they shown that title-based

document descriptors positively affected the performance of categorization.

Mani (88) made a distinction among different kinds of summaries: an *extract* consists entirely of material copied from the input; an *abstract* contains material that is not present in the input, or at least expresses it in a different way; an *indicative abstract* is aimed at providing a basis for selecting documents for closer study of the full text; an *informative abstract* covers all the salient information in the source at some level of detail; and a *critical abstract* evaluates the subject matter of the source document, expressing the abstractor views on the quality of the author's work.

According to (71), summarization techniques can be divided in two groups: those that extract information from the source documents (*extraction-based approaches*) and those that abstract from the source documents (*abstraction-based approaches*). The former impose the constraint that a summary uses only components extracted from the source document. These approaches put strong emphasis on the form, aiming at producing a grammatical summary, which usually requires advanced language generation techniques. The latter relax the constraints on how the summary is created. These approaches are mainly concerned with what the summary content should be, usually relying solely on extraction of sentences.

Although potentially more powerful, abstraction-based approaches have been far less popular than their extraction-based counterparts, mainly because generating the latter is easier. Some of the most effective techniques in text retrieval and classification rely on the bag of words representation, for which the document is represented as an unordered set of terms. An extraction-based summary consists of a subset of words from the original document and its bag of words (BoW) representation can be created by selectively removing a number of features from the original term set. In text classification this approach is known as "feature selection", that relies on the usefulness of each term as far as the classification accuracy is concerned. In the Text Summarization context this aspect is not relevant. Typically, an extraction-based summary whose length is only 10-15% of the original is likely to lead to a significant feature reduction as well.

One may argue that extraction-based approaches are too simple. However, as shown in (28), extraction-based summaries of news articles can be more informative than those resulting from more complex approaches. Also, headline-based article descriptors proved to be effective in determining user's interests (70).

Classic Extractive Techniques

In the work of Kolcz et al. seven straightforward (but effective) extraction-based text summarization techniques have been proposed and compared. In all cases, a word occurring at least three times in the body of a document is a keyword, while a word occurring at least once in the title of a document is a title-word.

For the sake of completeness, let us recall the proposed techniques:

- *Title* (T), the title of a document;
- *First Paragraph* (FP), the first paragraph of a document;
- *First Two Paragraphs* (F2P), the first two paragraphs of a document;
- *First and Last Paragraphs* (FLP), the first and the last paragraphs of a document;
- *Paragraph with most keywords* (MK), the paragraph that has the highest number of keywords;
- *Paragraph with most title-words* (MT), the paragraph that has the highest number of title-words;
- *Best Sentence* (BS), sentences in the document that contain at least 3 title-words and at least 4 keywords.

The effectiveness of these techniques could be attributed to the fact that a document (e.g., a scientific paper) is usually written to capture the user's attention whit the headlines and initial text (e.g., the abstract or the introduction). The last part of a document could also contain relevant content (e.g., the conclusions).

As the input of a contextual advertiser is an HTML document, CA systems typically rely on extraction-based approaches, which are applied to the relevant blocks of a webpage (e.g., the title of the webpage, its first paragraph, and the paragraph which has the highest title-word count).

4.2 Development and Evaluation of Novel Techniques

The methods introduced by Kolcz were introduced and evaluated for conventional documents (e.g., books, news, scientific papers). Since in CA the input is an HTML code the classic methods could be less effective. In fact, a webpage is usually more concise and noisy than a textual document, and it is not often written in "standard English". In fact, on the one hand a webpage is shorter than a paper, for example, and could

contain different items that lie outside from the body content (such as, metadata or anchor text). On the other hand, the frequency of slang, typos, and non conventional textual elements (e.g., emoticons or abbreviations) is higher than in textual documents, special in blogs, social networks, personal pages, or some other user's generated page.

To improving the performances of the Kolcz's methods we looked for some additional features to be used in conjunction with the classic features. The selected feature has been the title of the webpage (usually wrapped by the "title" tag). We remark that differently by Kolcz's technique, the title is not the headline of a textual section (contained for instance in the "h1" tags). Our proposal consists of the following techniques (14):

- *Title and First Paragraph* (TFP), the title of a document and its first paragraph:
- *Title and First Two Paragraphs* (TF2P), the title of a document and its first two paragraphs;
- *Title, First and Last Paragraphs* (TFLP), the title of a document and its first and last paragraphs;
- *Most Title-words and Keywords* (MTK), the paragraph with the highest number of title-words and that with the highest number of keywords.

We also defined a further technique, called *NKeywords* (NK), that selects the N most frequent keywords.[1].

4.2.1 Experimental Results

We performed experiments aimed at comparing the techniques described in Section 4.2.

To perform comparative experiments, we devised a suitable system, which is depicted in Figure 4.1, in which the *Text Summarizer* is the module aimed at performing text summarization and the *Classifier* is a centroid-based classifier adopted to classify each page in order to calculate precision, recall, and $F_{measure}$ of the adopted text summarization techniques. In other words, to evaluate the text summarization techniques, we used a Rocchio classifier[2] (114) with only positive examples and no relevance feedback, preliminary trained with about 100 webpage for class. Thus, pages are classified by considering the highest score(s) obtained by the cosine similarity method. In

[1] N is a global parameter that can be set starting from some relevant characteristics of the input (e.g., from the average document length).

[2] The Rocchio classifier is described in the next Chapter

order to evaluate the effectiveness of the classifier, we performed also a preliminary experiment in which pages are classified without relying on text summarization. The classifier shown a precision of 0.862 and a recall of 0.858. The classifier has the same implementation for the one used for extracting the classification features described in Section 3.1

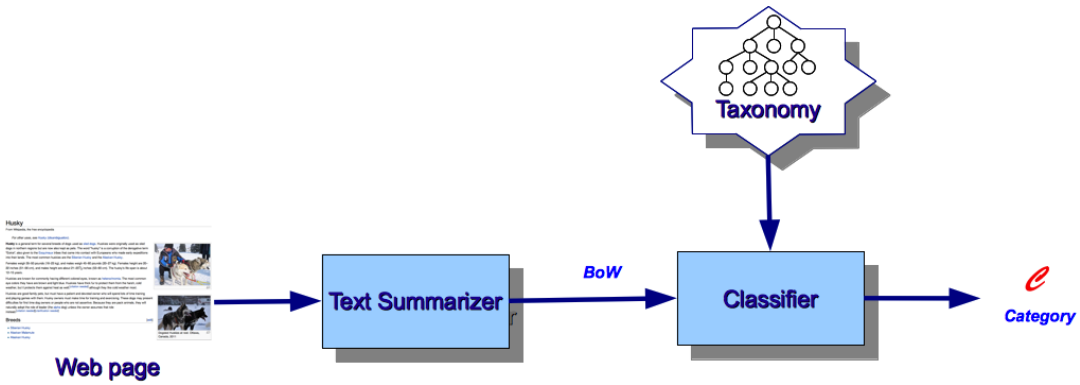

Figure 4.1: The system adopted to perform comparative experiments on text summarization.

Then, we performed comparative experiments among the methods of Kolcz et al. (see Section 4.1), except "Best Sentence" [1] and the corresponding enriched TS techniques proposed in 4.2.

Table 4.1 shows the performances in terms of macro-precision (P), macro-recall (R), and F-measure (F1). For each technique, the average number of unique extracted terms (T) is shown. The results show that just adding information about the title improves the performances of TS. Another interesting result is that, as expected, the TFLP summarization provides the best performance, as FLP summarization does for the classic techniques.

4.3 Using Snippets in Text Summarization: a Comparative Study and an Application

The problem of automated Text Summarization nowadays is in part due to the dynamic content of a webpage. In fact, classic techniques are no more available for Web created

[1] This method was defined to extract summaries from textual documents such as articles, scientific papers and books. In fact, we are interested in summarizing HTML documents, which are often too short to find meaningful sentences composed by at least 3 title-words and 4 keywords in the same sentence.

Table 4.1: Comparative results on TS techniques.

	P	R	F1	T
T	**0.798**	0.692	0.729	3
FP	0.606	0.581	0.593	13
F2P	0.699	0.673	0.686	24
FLP	0.745	**0.719**	**0.732**	24
MK	0.702	0.587	0.639	25
MT	0.717	0.568	0.634	15
TFP	0.802	0.772	0.787	16
TF2P	0.822	0.789	0.805	27
TFLP	**0.832**	**0.801**	**0.816**	26
MTK	0.766	0.699	0.731	34

by the adoption of tools for dynamic generation, such as Microsoft Silverlight[1], Adobe Flash[2], Adobe Shockwave[3], or for pages that contain applets written in Java. So we propose a novel method that relies on the adoption of snippets (i.e., the page excerpts provided by the search engines following the user's query), able, at least in principle, to give a relevant content of the suggested links in few lines and features. We claim that a snippet could be used as to perform Text Summarization (17).

4.3.1 Snippets in Search Engines

A general definition of *snippet* is "a small piece of something". In programming, it refers to a small region of re-usable source code, machine code, or text. Snippets are often used to clarify the meaning of an otherwise cluttered function, or to minimize the use of repeated code that is common to other functions.

Snippets are also used by search engines to provide an excerpt of the corresponding webpage according to the keywords used in the query. Snippet can be considered as a topic-driven summarization, since the summary content depends on the preferences of the user and can be assessed via a query, making the final summary focused on a particular topic. In replying to a user's query, search engines provide a ranked list of related webpages, each described by a title, a set of snippets, and its URL (see Figure 4.2). The tile is directly taken from the *title* tag of the page, whereas the URL is the

[1]http://www.microsoft.com/silverlight/
[2]http://www.adobe.com/products/flashplayer.html
[3]http://get.adobe.com/it/shockwave/

http address of the page.

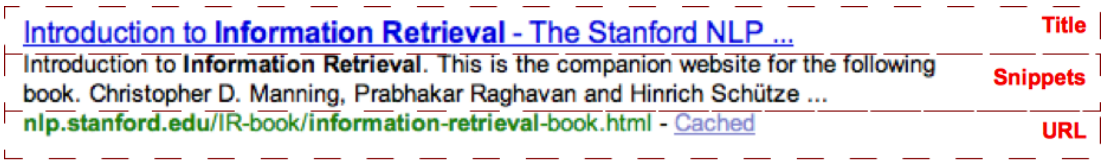

Figure 4.2: An example of results given by Yahoo! search engine for the query "Information retrieval".

The choice of a snippet, for a search engine, is an important task. If a snippet shown to the user is not very informative, the user may click on pages in search results that do not contain the information s/he is looking for, or s/he may not click on pages that may be helpful. Moreover, poorly chosen snippets can lead to bad searching experiences. Snippets are usually directly taken from the *description meta* tag, if available. If the description meta tag is not provided, the search engine may use the description for the site provided by the Open Directory Project (aka, DMoz)[1], or a summary extracted from the main content of the page.

Snippet extraction depends on the adopted search engine. Google[2] not always uses the meta description of the page. In fact, if the content provided by the Web developer in the description meta tag is not helpful, or less than reasonable quality, then Google replaces it with its own description of the site. In so doing, Google snippet will be different depending on the user's search query. Yahoo![3] provides a patent application that describes how to better decide which snippet to show to users. The gist of Yahoo! patent application is based on three main issues[4]: (i) a query-independent relevance for each line of text, i.e., a degree to which the line of text of the document summarizes the document; (ii) a query-dependent relevance of each of the lines of text, i.e., a relevance of the line of text to the query; and (iii) the intent behind a query. To our best knowledge, Bing[5] developers do not give information on how snippets are extracted.

4.3.2 Experimental Results

To evaluate the method we adopt the same classifier used for the evaluation of the techniques described in section 4.2.

[1] http://dmoz.org
[2] http://www.google.com
[3] http://www.yahoo.com
[4] http://www.seobythesea.com/2009/12/how-a-search-engine-may-choose-search-snippets/
[5] http://www.bing.com

Experiments Set Up

Experiments have been performed on the Recreation and BankSearch datasets, described in Section 3.2.

As a baseline for our comparative experiments, we adopt the text summarization technique called $TFLP$ (Title, First and Last Paragraph summarization), described in the previous section. That technique, as we showed in the previous Section, showed the best results compared with the state-of-the-art techniques proposed by Kolcz. As for the snippets, we performed a query to Yahoo! and we used the returned snippets. We performed experiments by considering the snippets by themselves (S) and in conjunction with the title of the corresponding webpage (ST). It is worth noting that we disregarded dynamic pages from both datasets in order to process the same number of pages independently by the adopted text summarization technique, allowing a fair comparison.

Results

Table 4.2 reports our experimental results in terms of precision (P), recall (R), and $F_{measure}$ (F_1). The Table gives also the average number of extracted terms (T).

Table 4.2: Results of text summarization techniques comparison.

	BankSearch			**Recreation**		
	TFLP	**S**	**ST**	**TFLP**	**S**	**ST**
P	**0.849**	0,734	0.806	0.575	0.544	**0.595**
R	**0.845**	0.730	0.804	**0.556**	0.506	0.554
F1	**0.847**	0.732	0.805	0.565	0.524	**0.574**
T	26	12	14	26	11	13

Results show that all the adopted techniques obtained better results in BankSearch than in Recreation. Moreover, they point out that, in both datasets, results obtained by relying on snippets together with the title (ST) are comparable with those obtained by adopting $TFLP$. In particular, $TFLP$ performs slightly better in BankSearch, whereas ST performs slightly better in Recreation. This proves that snippets can be adopted as text summarization techniques, especially when classical techniques can not be applied, as in the case of dynamic webpages.

4.4 Text Summarization in Contextual Advertising

Since we are mainly interested in studying the impact of the syntactic phase on CA, we implemented a CA system for each technique, and we evaluated the performances by comparative experiments.

4.4.1 The Impact of the Enriched Techniques

To study the impact of the techniques described in Section 4.2 in CA, we implemented the system described in Section 3.1, comparing results while varying the adopted TS technique (16). Let us note that, even if modern advertising networks should work in real time, for the sake of completeness we calculated performances without exploiting Text Summarization (the first row in Table 4.3).

Table 4.3: Comparative results on CA.

	P@1	α_b	**P@2**	α_b	**P@3**	α_b	**P@4**	α_b	**P@5**	α_b
no TS	**0.785**	0.0	**0.779**	0.0	**0.775**	0.0	**0.769**	0.0	**0.753**	0.2
T	**0.680**	0.1	**0.675**	0.1	**0.652**	0.1	**0.639**	0.1	**0.595**	0.3
FP	0.488	0.2	0.476	0.3	0.448	0.3	0.421	0.2	0.391	0.1
F2P	0.613	0.2	0.609	0.2	0.588	0.1	0.558	0.2	0.514	0.3
FLP	0.674	0.0	0.653	0.0	0.617	0.2	0.582	0.2	0.546	0.1
MK	0.631	0.2	0.620	0.1	0.581	0.1	0.541	0.2	0.500	0.3
MT	0.640	0.4	0.617	0.3	0.610	0.2	0.586	0.1	0.547	0.3
TFP	0.744	0.0	0.707	0.3	0.691	0.1	0.669	0.0	0.637	0.1
TF2P	0.740	0.0	0.723	0.1	0.721	0.1	**0.712**	0.1	**0.678**	0.0
TFLP	**0.768**	0.2	**0.750**	0.2	**0.729**	0.3	0.701	0.3	0.663	0.0
MTK	0.711	0.2	0.698	0.1	0.685	0.2	0.663	0.2	0.608	0.2

To evaluate the performances we adopt the BankSearch dataset described in Section 3.2. Also we manually built a repository of ads, in which each ad is represented by the webpage code, processed in the same way as for the input page (target page). We selected 5 relevant ads for each category, so we have 55 total ads.

CA systems choose the relevant ads contained in the repository according to the scores obtained by the Matcher. The ads with the highest scores are displayed on the target page.

Five different experiments have been performed for each system, in which from 1 to 5 ads are selected for the target page, respectively. Table 4.3 reports, for each TS

technique, the precision at k ($k = [1, ...5]$) in correspondence of the best value of α (α_b). we implemented the system described in Section 3.1.

As expected, the best results are obtained without adopting any Text Summarization technique. Among the selected techniques, TFLP has the best performances in terms of $P@1$, $P@2$, and $P@3$, whereas TF2P in terms of $P@4$ and $P@5$.

Table 4.4: Results with TFLP by varying α.

α	**p@1**	**p@2**	**p@3**	**p@4**	**p@5**
0.0	0.765	0.746	0.719	0.696	**0.663**
0.1	0.767	0.749	0.724	0.698	**0.663**
0.2	**0.768**	**0.750**	**0.729**	0.699	0.662
0.3	0.766	0.749	**0.729**	**0.701**	0.661
0.4	0.756	0.747	**0.729**	0.698	0.658
0.5	0.744	0.735	0.721	0.693	0.651
0.6	0.722	0.717	0.703	0.681	0.640
0.7	0.685	0.687	0.680	0.658	0.625
0.8	0.632	0.637	0.635	0.614	0.586
0.9	0.557	0.552	0.548	0.534	0.512
1.0	0.408	0.421	0.372	0.388	0.640

To further highlight the impact of TS, Table 4.4 reports the results obtained by using TFLP while varying α for the p@1. According to equation (3.1), a value of 0.0 means that only semantic analysis is considered, whereas a value of 1.0 considers only the syntactic analysis. A comparative study of the role of the α parameter is reported in Figure 4.3, in which the behavior of $P@1$ by varying α is reported for each classic and novel technique. In this figure is more clear how the best technique is the TFLP.

4.4.2 The Impact of Snippets

Being interested in studying the impact of snippets, we devised a suitable system (see Figure 4.4). The system takes a webpage as input. The *BoW builder*, first, performs queries to Yahoo!, asking for the URL of each webpage of the dataset, and uses the returned snippets. Then, stop-words are removed and a stemming task is performed. This module outputs a vector representation of the original text as BoW, each word being represented by its TFIDF. Starting from the BoW provided by the *BoW builder*, the *Classifier* classifies the page according to the given taxonomy by adopting a centroid-based approach. This module outputs a vector representation in terms of Classification

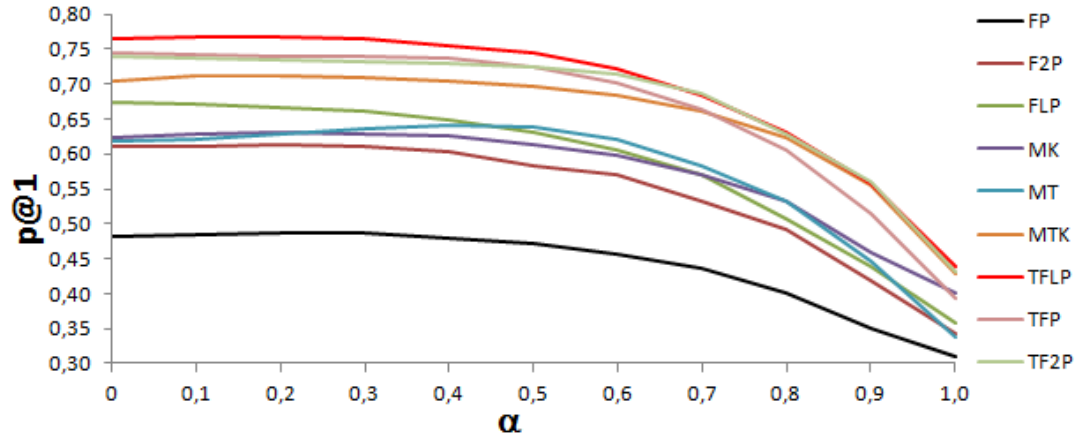

Figure 4.3: The behavior of p@1 by varying α.

Features (CF), each features corresponding to the score given by the classifier for each category. Finally, the *Matcher* ranks the categories according to the scores given by the classifier (i.e., the CF of the target page) and, for each category, randomly extracts from the *Ads repository* a corresponding ad.

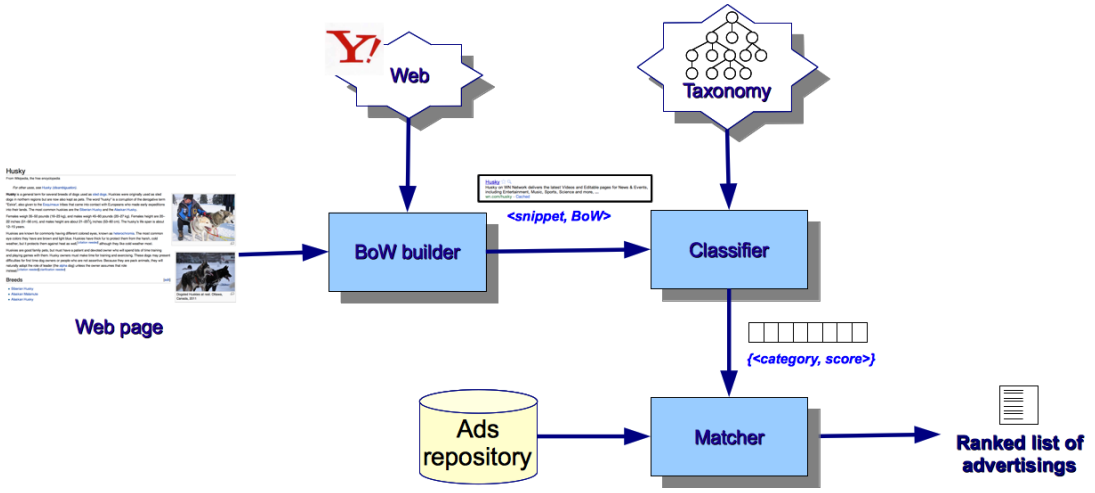

Figure 4.4: The implemented baseline system.

Let us note that the proposed system, except for the adopted Text Summarization technique, is compliant with the baseline system proposed in Chapter 2 in which only CF are considered in the matching phase.

System Performances

To assess the effectiveness of the proposed approach, experiments have been performed on the Recreation dataset. As for the ads to be suggested, we adopted the manually built ad repository described in the previous sections. In that repository, each ad is represented by the webpage of a product or service company.

Performances have been calculated in terms of *precision at k* with $k \in [1, 5]$, i.e., the precision in suggesting k ads.

Table 4.5: Precision at k of the proposed CA system by adopting: $TFLP$ (CA_{TFLP}), the sole snippets (CA_S); and the snippets together with the page title (CA_{ST}).

k	**CA**$_{TFLP}$	**CA**$_S$	**CA**$_{ST}$
1	**0.868**	0.837	0.866
2	0.835	0.801	**0.836**
3	0.770	0.746	**0.775**
4	0.722	0.701	**0.729**
5	0.674	0.657	**0.681**

In performing experiments, we compared the performances obtained by using as text summarization technique: $TFLP$, the resulting system being CA_{TFLP}; the sole snippets, the resulting system being CA_S; and the snippets together with the page title, the resulting system being CA_{ST}. Let us recall that, since the focus of this Chapter is on text summarization, comparative experiments among the implemented CA system and selected state-of-the-art systems are out of the scope of this work. Nevertheless, let us stress the fact that CA_{TFLP} coincides with the system proposed in Section 4.4.1 in which the α parameter is set to 0 (i.e., only CF are considered in the matching phase).

Table 4.5 shows that, for all the compared systems, the results are quite good, especially in suggesting 1 or 2 ads. It also clearly shows that, except for $k = 1$, CA_{ST} is the system that performs better. This proves the effectiveness of the adoption of snippets as text summarization technique in the field of CA.

4. THE IMPACT OF TEXT SUMMARIZATION ON CONTEXTUAL ADVERTISING

Chapter 5

The Impact of Semantics on Contextual Advertising

As shown in section 2.3.1, the approaches based solely on syntactic analyses of the original text can lead to poor performances of ad networks. In order to solve these problems, the state-of-the-art ad networks rely on approaches that combine syntactic and semantic analyses. The semantic analysis is based on the adoption of external lexical resources (typically, a taxonomy of topics), which is devoted to classify pages and ads to extract suitable features to be used in conjunction with the classic syntactic keywords. This chapter is focused on the importance of semantics, and a novel approach based on ConceptNet is proposed.

G.Armano, **A.Giuliani**, E.Vargiu. *Studying the Impact of Text Summarization on Contextual Advertising*, KDIR'11: Proceedings of International Conference on Knowledge Discovery and Information Retrieval, 2011.

5.1 Semantic Classification

In state of the art approaches (as shown in Figure 3.1) the process of selecting the suitable ads is based on two distinct analyzes, syntactic and semantic. In this section we are interested at the semantic phase.

The semantic phase relies on a classification of both page and ad into taxonomy of topics, in order to find the proximity of the page and the ad classes, and to extract semantic features (called Classification Features). The target is to favor ads topically related to the page, and the classification features give rise to the most related topics. Furthermore, the adoption of a hierarchical taxonomy could allow for a gradual generalization of the ad search space if there are no specific ads for the precise topic of the page.

Broder et al. claimed to solve a challenging problem: classify both page and ads within a large taxonomy (to have a small topic granularity) with high precision. They proved that the best classifier to extract discriminating classification features is a variation of the Rocchio classifier, which is described in the following section.

5.1.1 Rocchio Classification for Text Categorization

Rocchio Classification(114) is based on a method of relevance feedback developed for IR systems around the year 1970. Like many other retrieval systems, the Rocchio feedback approach was developed using a Vector Space Model. The algorithm is based on the assumption that most users have a general conception of which documents should be denoted as relevant or non-relevant. Therefore, the user's search query is revised to include an arbitrary percentage of relevant and non-relevant documents as a means of increasing the search engine's recall, and possibly the precision as well.

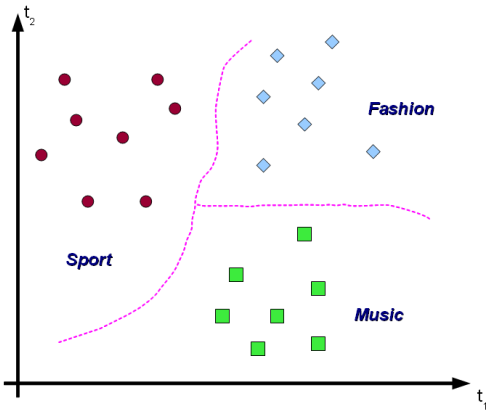

Figure 5.1: Example of vector space model for textual documents.

This kind of classification can be adapted for text categorization; relevance feedback could be viewed as a binary classification (relevant and non-relevant documents). To this end, we consider the feature space as the space obtained taking into account the

meaningful terms of a set of documents, in which each dimension of the space corresponds to a distinct term. The adapted Rocchio classification relies only in positive examples and no relevance feedback, as proposed by Broder.

For instance, Figure 5.1 shows three classes, *Sport*, *Fashion*, and *Music*, in a two-dimensional space. The axis represent the weights of each meaningful term (t_1 and t_2). For each class, documents are shown as circles, diamonds, and squares respectively. The first step is to find a good separator for each class; an example is depicted in the same figure, in which the classes are separated by the dotted lines. The dotted lines are called *decision boundaries*.

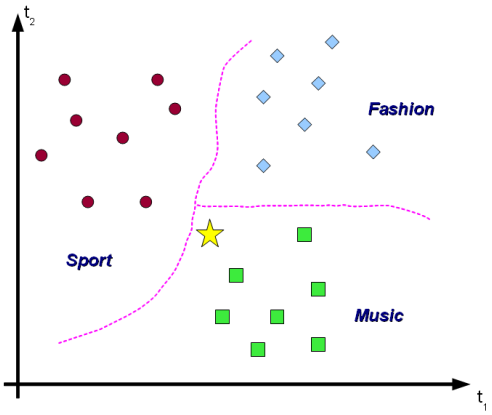

Figure 5.2: Classification of a new document.

To classify a new document, depicted as a star in Figure 5.2, the region in which it occurs is chosen as the assigned class (*Music* in the example). The task in vector space classification is to devise algorithms that compute the best boundaries, in term of high accuracy on test data. One of the best way to compute good boundaries, and the most used in IR, is the *Rocchio classification*, which uses *centroids* as boundaries.

The centroid μ of a class c is computed as the average vector (or center of mass) of its members:

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{\vec{d} \in D_c} \frac{\vec{d}}{\left\|\vec{d}\right\|} \tag{5.1}$$

in which D_c is the number of proper documents for the class c and d is a generic document. For the previous example, the Figure 5.3 shows the centroids (the dark crosses) for the three classes.

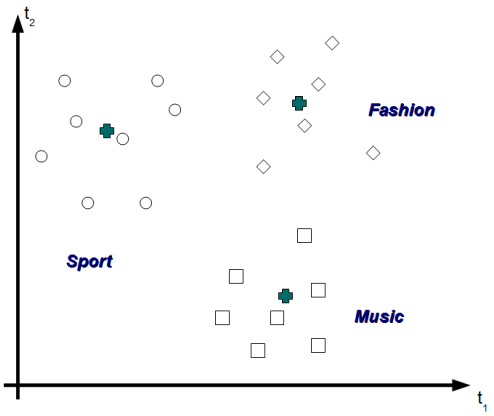

Figure 5.3: Centroids of classes.

The boundary between two classes in Rocchio classification is the set of points with equal distance from the two centroids. For the three sample classes, the following figure depicts the boundaries of separation:

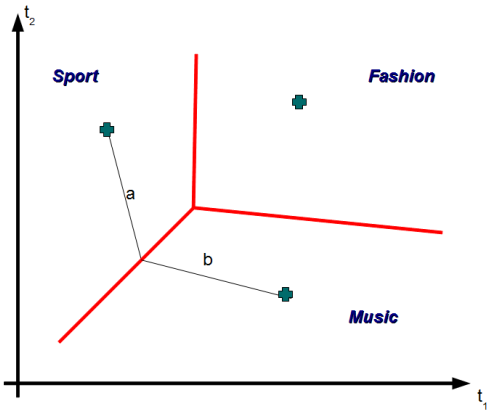

Figure 5.4: Boundaries in Rocchio classification.

If we take a point of a line, the distances from the two centroids of adjoining classes (Sport and Music in the example) have the same value ($|a| = |b|$).

If we extend all the previous issues to a multidimensional space, the boundaries are hyperplanes that divide the space in distinct regions, and the classification task is performed by selecting for a new document the class belonging to the region in which the document falls. Equivalently, the task is performed by measuring the distances between the document and each centroid, and selecting the class the centroid with the

minimum distance belongs to.

5.1.2 Rocchio Classifier for Contextual Advertising

To infer the topics of each ad and of the original page, they are classified according to a given taxonomy. First, for each node of the taxonomy all its documents are merged into a single compound document. Such document is then used as a centroid for the Rocchio classifier described in the previous section. Each centroid is defined as a sum of the TFIDF values of each term, normalized by the number of training documents in the class. In formula:

$$\vec{c_j} = \frac{1}{|C_j|} \sum_{\vec{d} \in C_j} \frac{\vec{d}}{\left\|\vec{d}\right\|} \tag{5.2}$$

where $\vec{c_j}$ is the centroid for class C_j and d ranges over the documents of a particular class. The classification of a document (page or ad) is based on the cosine of the angle between the document d and the centroid of the class C_j; in formula:

$$C^* = \underset{c_j \in C}{argmax} \left(\frac{\vec{c_j}}{\left\|\vec{c_j}\right\|} \cdot \frac{\vec{d}}{\left\|\vec{d}\right\|} \right) = \underset{c_j \in C}{argmax} \frac{\sum_{i \in |F|} c_j^i \cdot d^i}{\sqrt{\sum_{i \in |F|} \left(c_j^i\right)^2} \sqrt{\sum_{i \in |F|} \left(d^i\right)^2}} \tag{5.3}$$

where F is the set of features. To produce comparable scores, each score is normalized with the document and the class length. The terms c_j^i and d^i represent the weight of the ith feature, based on the standard TFIDF formula, in the class centroid and the document, respectively.

5.2 Semantic Enrichment of Contextual Advertising by Using Concepts

Since Broder proved that semantic information improves the performances if it is taken in conjunction with syntactic features. The goal of this part of the research activity has been focused on the study further semantic information, and the development of a system in order to evaluate it. The very novel contribution of the proposed system has been the adoption of concepts to semantically enrich the content matching. In

particular, to infer the concepts we resort to ConceptNet. ConceptNet is a semantic network automatically built by data collected in the Open Mind Common Sense (OMCS) project.

5.2.1 ConceptNet

The Open Mind Common Sense (OMCS) project is a distributed solution to the problem of commonsense acquisition by enabling users to enter commonsense into the system with no special training or knowledge of computer science. In 2000, the OMCS project began to collect statements from untrained volunteers on the Internet. These data have been used to automatically build a semantic network, called ConceptNet (83).

In ConceptNet, nodes are concepts and edges are predicates. Concepts are aspects of the world that people would talk about in natural language. They correspond to selected constituents of the commonsense statements that users have entered, and can represent *noun phrases*, *verb phrases*, *adjective phrases*, or *prepositional phrases*. Predicates express relationships between two concepts. They are extracted from natural language statements enter by contributors, and express relationships such as *IsA*, *PartOf*, *LocationOf*, and *UserFor*. In addition, there are also some underspecified relation types such as *ConceptuallyRelatedTo*, which means that a relationship exists between two concepts without any other semantic explanation.

Predicates in ConceptNet are created by a pattern-matching process. Each sentence is compared with an ordered list of patterns, which are regular expressions that can also include additional constraints on phrase types based on the output of a natural language tagger and chunker. These patterns represent sentence structures that are commonly used to express different relationships. The phrases that fill the slots in a pattern will be turned into concepts. When a sentence is matched against a pattern, the result is a "raw predicate" that relates two strings of text. A normalization process determines which two concepts these strings correspond to, turning the raw predicate into an edge of ConceptNet. In Figure 5.5 a sample of ConceptNet is sketched.

ConceptNet has been adopted in several application fields. To our best knowledge this is the first attempt to use ConceptNet in CA.

In our system we use the the ConceptNet 3 database (58) to find the related meaningful concepts. Each element of the database is called *assertion*, and is uniquely defined by several attributes, for instance: (i) a couple of concepts, i.e., two linked nodes of ConceptNet; (ii) the language of the concepts, e.g., English, Italian, or French; (iii) the type of relation that connects the two concepts, e.g., "IsA" or "PartOf"; (iv) the score, i.e., a value given by the users that represents the reliability of the assertion; and (v)

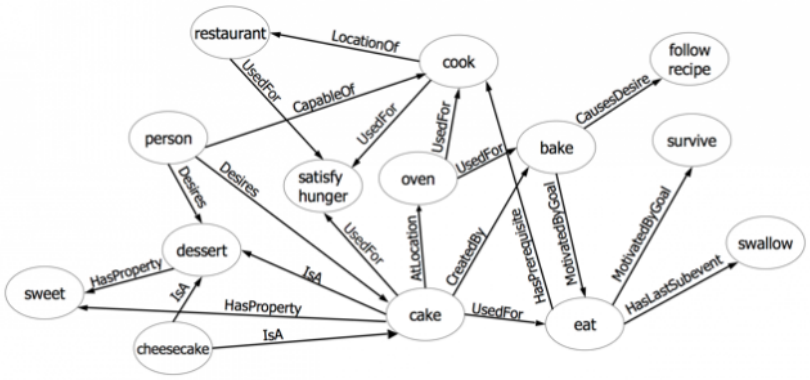

Figure 5.5: A sample of ConceptNet.

the frequency, i.e., a textual value (ranging from "never" to "always") that expresses how often a relation connects a given couples of concepts.

5.2.2 ConCA: Concepts in Contextual Advertising

ConCA has been implemented in Java and, as shown in Figure 5.6, it encompasses four main modules: (i) *Text Summarizer*; (ii) *Classifier*; (iii) *Concept Extractor*; and (iv) *Matcher* (15).

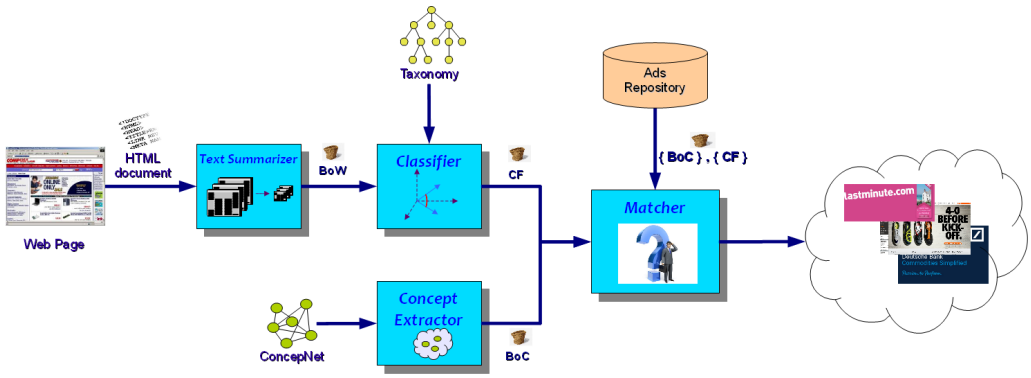

Figure 5.6: ConCA architecture.

Text Summarizer

It outputs a vector representation of the original HTML document as Bag of Words (BoW), each word being weighted by TF-IDF. We adopted the TFLP technique (see Section 5.1), which considers information belonging to the Title, the First, and the Last Paragraph of the Web document (the page or the ad).

Classifier

It takes in input the BoW provided by the Text Summarizer and outputs the set of Classification Features (CF) inferred by the centroid-based classifier. The classification features are the scores of the similarities between the page and each class centroid, as we showed in the previous section.

Concept Extractor

It takes the BoWrepresentation as input and, for each term, queries the ConceptNet 3 database in order to obtain the set of assertions that include the term, each being one of the two concepts in the assertion. The resulting assertion set is then filtered in order to reduce noise. In particular, we consider only the assertions that satisfy the following constraints:

- the language of the concepts is English;
- the type of the relation is "IsA" or "HasA";
- the score given by users is greater than 1;
- the frequency is "positive", e.g., assertions with "never" are not considered.

As for concepts, only those with a number of assertions greater than 3 are selected. The output is a set of concepts with their occurrences (calculated by TF-IDF), called Bag of Concepts (BoC).

The Concept Extractor model is depicted in Figure 5.7

- ***ConceptNet DB*** The ConceptNet 3.0 is stored in a suitable relational database. It receives a query and outputs the resulting concepts satisfying the query parameters.
- ***Concept Selector*** Given the bag of words, for each terms it queries the DB and retrieves the entire list of concepts associated to the testing page.

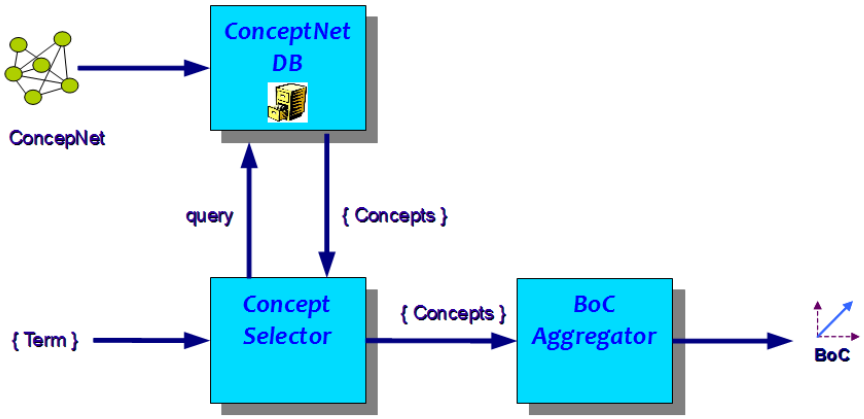

Figure 5.7: Classifier.

- ***BoC Aggregator*** It is devoted to the bag of concepts representation. It outputs a vector in which features are the associated weight of each term. The weight is calculated with the TFIDF of each concept.

Matcher

It is devoted to choose the relevant ads according to a score based on the similarity between the target page and each ad. The score is computed by the cosine similarity, i.e., the cosine of the angle between the two corresponding vectors:

$$sim(p, a_j) = \frac{\vec{p} \cdot \vec{a_j}}{|\vec{p}| \times |\vec{a_j}|} = \frac{\sum_{i=1}^{n} w_{ip} \cdot w_{ij}}{\sqrt{\sum_{i=1}^{n} w_{ip}^2} \sqrt{\sum_{i=1}^{n} w_{ij}^2}} \tag{5.4}$$

Being interested in studying the influence of concepts, we calculate a score similarity taking into account BoC and CF:

$$\sigma(p, a) = \alpha \cdot sim_{BoC}(p, a) + (1 - \alpha) \cdot sim_{CF}(p, a) \tag{5.5}$$

in which α is a global parameter that permits to control the impact of BoC with respect to CF, whereas $sim_{BoC}(p, a)$ and $sim_{CF}(p, a)$ are cosine similarity scores between p and a using BoC and CF, respectively.

Table 5.1: Results of CA systems comparison.

	Baseline System					**ConCA**				
α	**π@1**	**π@2**	**π@3**	**π@4**	**π@5**	**π@1**	**π@2**	**π@3**	**π@4**	**π@5**
0.0	0.765	0.746	0.719	0.696	**0.663**	0.765	0.746	0.719	0.696	0.663
0.1	0.767	0.749	0.724	0.698	**0.663**	**0.773**	**0.752**	**0.728**	**0.701**	**0.668**
0.2	**0.768**	**0.750**	**0.729**	0.699	0.662	0.761	0.747	0.724	0.696	0.662
0.3	0.766	0.749	**0.729**	**0.701**	0.661	0.736	0.730	0.709	0.685	0.650
0.4	0.756	0.747	0.728	0.698	0.658	0.701	0.704	0.686	0.668	0.636
0.5	0.744	0.734	0.720	0.692	0.651	0.661	0.664	0.660	0.643	0.614
0.6	0.722	0.717	0.703	0.681	0.640	0.609	0.624	0.623	0.614	0.588
0.7	0.684	0.687	0.680	0.658	0.625	0.561	0.573	0.578	0.568	0.551
0.8	0.632	0.637	0.635	0.614	0.586	0.512	0.518	0.517	0.513	0.501
0.9	0.557	0.552	0.548	0.534	0.512	0.481	0.471	0.462	0.455	0.440
1.0	0.439	0.421	0.408	0.388	0.372	0.427	0.407	0.394	0.379	0.360

5.2.3 Experiments and Results

To perform experiments we used the BankSearch Dataset, and a repository of 55 ads. both datasets are described in section 4.4.

To evaluate the performances of ConCA, our baseline is the system described in Section 4.4.

Five different experiments have been performed for each system augmenting the suggested ads from 1 to 5. Results are then calculated in terms of the precision $\pi@k$, in which k is the number of suggested ads in the range $[1, 5]$. In particular, we calculated micro-precision as follows:

$$\pi^{\mu} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m}(TP_i + FP_i)} \tag{5.6}$$

where TP_i (True Positives) are the retrieved ads of class i suggested to a page p of class i; FP_i (False Positives) are the retrieved ads that does not belong to class i suggested to a page p of class i; and m is the total number of classes.

For each experiment, Table 5.1 reports the results obtained by varying α. According to equation (5.5) a value of 0.0 means that only CF are considered, whereas a value of 1.0 considers only BoC in ConCA and BoW in the baseline system.

Results show that ConCA performs slightly better than the baseline system. In particular, for both systems, the best performance is obtained with a low value of α (i.e., 0.1 for ConCA). It means that CF have more impact than BoC in ConCA. Similarly, CF have more impact than BoW in the baseline system. Nevertheless, concepts positively affect results when suggesting 1 or 5 ads. Since these preliminary results

are encouraging, we are currently performing experiments that consider BoW, BoC, and CF in conjunction. In this way, we adopt both the syntactic and the semantic contributions.

As a final remark, running times for ConCA and the baseline system are comparable, so that the real-time constraint is preserved.

Chapter 6

Collaborative Filtering on Contextual Advertising

CA can be viewed as an information filtering task aimed at selecting suitable ads to be suggested to the final "user", i.e., the webpage in hand.

Web 2.0 users need suggestions about online contents (e.g., news and photos), people (e.g., friends in social networks), goods for sale (e.g., books and CDs), and/or services and products (e.g., suitable ads) according to their preferences and tastes. In this scenario, Information Filtering (IF) techniques, aimed at presenting only *relevant* information to users, requires improvements to make filtering methods more robust, intelligent, effective, and applicable to a broad range of real life applications. To this end, the research in this field is focused on defining and implementing intelligent techniques, which make use of machine learning, text categorization, evolutionary computation, and semantic web (13).

IF is typically performed by RS. In particular, recommendations are provided by relying on Collaborative Filtering (CF). CF consists of automatically making predictions (*filtering*) about the interests of a user by collecting preferences or tastes from similar users (*collaboration*); the underlying idea is that those who agreed in the past tend to agree again in the future.

Several CF systems have been developed to suggest items and goods, including news, photos, people, and books (6). To our best knowledge, CF has not been adopted yet to perform Web advertising, i.e., to suggest ads to a webpage. Since the task of suggesting an ad to a webpage can be viewed as the task of recommending an item (the ad) to a user (the webpage) (18), we claim that also Web advertising systems could be developed by exploiting CF.

A.Addis, G.Armano, **A.Giuliani**, E.Vargiu. *A Recommender System based on a Generic Contextual Advertising Approach*, Proceedings of ISCC'10: IEEE Symposium on Computers and Communications.

A. Addis, G. Armano, **A.Giuliani**, and E. Vargiu. *A Novel Recommender System Inspired by Contextual Advertising Approach*, Proceedings of IADIS'10: International Conference Intelligent Systems and Agents (ISA), 2010.

G. Armano, **A.Giuliani**, and E. Vargiu. *Intelligent Techniques in Recommendation Systems: Contextual Advancements and New Methods*, S. Dehuri, M.R. Patra, B.B. Misra, A.K. Jagadev (eds.), IGI Global (in press).

G.Armano, **A.Giuliani**, E.Vargiu. *Improving Contextual Advertising by Adopting Collaborative Filtering.* Transactions on the Web (submitted).

6.1 Recommender Systems

RS help users to navigate through large product assortments and to make decisions in e-commerce scenarios. The development of an RS is a multi-disciplinary effort, which involves experts from various fields –such as Artificial Intelligence, Human Computer Interaction, Information Technology, Data Mining, Statistics, Adaptive User Interfaces, Decision Support Systems, Marketing, and Consumer Behavior (112).

6.1.1 The Problem

The recommendation problem can be formulated as follows: let U be the set of all users[1] and let I be the set of all possible items that can be recommended (e.g., books, movies, and restaurants)[2]. Let f be a utility function that measures the usefulness of item i for user u, i.e., $f : U \times I \rightarrow R$, where R is a totally ordered set (e.g., non-negative integers or real numbers within a certain range). Then, for each user $u \in U$ we want to choose the item $i' \in I$ that maximizes the user's utility function. More formally:

$$\forall u \in U : i'_u = \underset{i \in I}{argmax} \;\; f(u, i) \tag{6.1}$$

In RS, the utility function is typically represented by ratings and is initially defined only on items previously rated by the users. For example, in a book recommendation

[1] It can be very large-millions in some cases.

[2] It can be very large, ranging in hundreds of thousands or even millions of items in some applications.

application (e.g., Amazon.com), users initially rate some subsets of books they have read.

6.1.2 Background

Based on how recommendations are made, RS fall in the following categories: content-based, collaborative filtering, and hybrid approaches.

Content-based RS suggest to users items similar to those they preferred in the past (84). Many current content-based systems focus on recommending items containing textual information, such as documents and Websites. The improvement over traditional IR approaches comes from the adoption of user profiles, which contain information about users' tastes, preferences, and needs. The profiling information can be elicited from users explicitly (e.g., through questionnaires) or implicitly learned from their transactional behavior over time.

Unlike content-based recommendation methods, collaborative RS –or collaborative filtering systems– try to predict the utility of items for a particular user based on the items previously rated by other users. There have been many collaborative systems developed in the academia and in the industry. Among others, let us recall: the Grundy system (113), GroupLens (110), Video Recommender (62), and Ringo (122) which have been the first systems that used CF algorithms to automate recommendation. Other examples of collaborative RS are: the book RS from Amazon.com (82) and the PHOAKS system that helps people to find relevant information on the Web (129).

The choice of the most suitable algorithm for a RS depends on many issues, including the specific type of service, the nature of items, together with the kind and amount of available information. For instance, if items are documents, an algorithm based on content matching is more appropriate, because it is able to deal with problems related to the automatic analysis of text (138) (128). If items are multimedia with scarce descriptions, but rated by a community of users, CF could be more suitable (119) (82).

Several RS use a hybrid approach which combines collaborative and content-based methods (32). In so doing, certain limitations of content-based and collaborative systems can be overcome. Different ways for giving rise to hybrid RS while combining collaborative and content-based methods have been proposed (6).

To support the recommendation process in social activities, group RS have also been proposed. These systems are aimed at providing recommendations to groups instead of individuals (27). The way a group is formed affects the way it is modeled and how recommendations are predicted. Four notions of group have been defined (64): (i) established group, i.e., a number of persons who explicitly choose to be a part of a

group; (ii) occasional group, i.e., a number of persons who occasionally do something together; (iii) random group, i.e., a number of persons who share an environment in a given moment without explicit interests that link them; and (iv) automatically identified group, i.e., groups that are automatically detected considering the preferences of the user and/or the available resources.

More recently, some works focused on combining information about user profiles with context information (7) (108). The underlying motivation is that taking into account both profiles and contexts in a recommendation process gives benefits to RS for many reasons: (i) user's preferences and ratings change according to their contexts (127); (ii) traditional RS do not consider multiple ratings of the same content (10); and (iii) RS may fail while trying to provide some valuable recommendations, as their similarity distance is uniformly applied to user's preferences without analyzing the discrepancies introduced by the context (1). In the literature, the term "context" is referred to *events which modify the user behavior* in the area of RS, and to *keywords used in search engines* in the area of CA. In this thesis, we always adhere to the latter interpretation. Therefore, we are not interested in context-aware RS.

6.2 Unifying view of CA and RS

As discussed in Section 2.2, CA is an interplay of four players:

- the *advertiser*, who provides ads and organizes her/his activity around campaigns, defined by a set of ads with a particular temporal and thematic goal (e.g., sale of digital cameras during the holiday season);
- the *publisher*, who is the owner of the webpages on which the advertising is displayed. Her/his goal is to maximize advertising revenue while providing a good user experience;
- he *ad network*, which, as a mediator between the advertiser and the publisher, selects the ads to display on the webpages;
- the *users*, who visit the webpages of the publisher and interact with the ads.

Figure 2.1 sketches the interactions among the four players in a typical CA task. Upon a request initiated by the users browser (HTTP get request), the Web server returns the requested page. As the page is being displayed, a JavaScript code embedded into the page (or loaded from a server) sends to the ad network a request for ads that contains the page URL and some additional data. The ad network model aligns the

interests of publisher, advertisers and the network itself. In general, user clicks bring benefits to the publisher and to the ad network by providing revenues, and to the advertiser by bringing traffic to the ad webpage.

Nothing prevents from viewing a RS as an interplay of four players (18), as depicted in Figure 6.1: (i) the *recommender*, who provides the supply of items to be recommended; (ii) the *publisher*, who is the owner of the webpages on which items are displayed for recommendation; (iii) the *recommender system*, which, as a mediator between recommender and publisher, is in charge of selecting the items to be recommended to a specific user; and (iv) the *user*, who visits the webpages of the publisher/recommender and interacts with the suggested items. Let us note that in RS the recommender and the publisher are typically the same player.

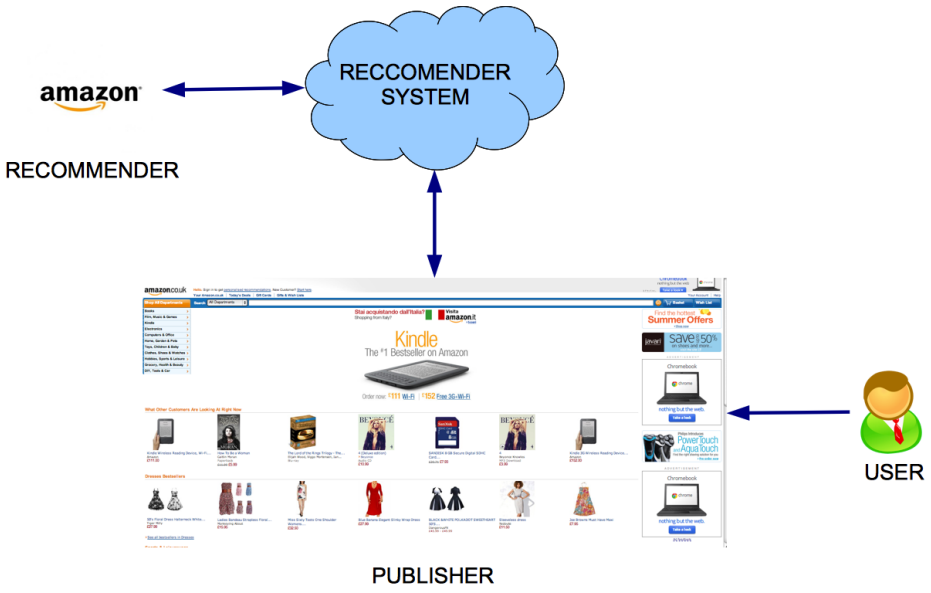

Figure 6.1: The four players in a RS task.

6.3 A Collaborative Filtering Approach to Perform Contextual Advertising

In the previous section we claim that a CA task can be viewed as a recommendation task. So that, we claim that CA systems could be improved by exploiting collaborative filtering through the extraction of suitable information from semantically related links.

In this Section, we present an experimental study aimed at investigating if related

links are effective for CA. To our best knowledge this is the first attempt to assess the effectiveness of semantically related links in the field of CA.

6.3.1 Semantically Related Links

The benefits of link information for IR have been well researched (73). Linked-based ranking algorithms use the implicit assumption that linked documents tend to be related each other and, therefore, that link information is potentially useful for retrieval and filtering (40) (68) (78) (121).

Marchiori (90) claims that "The power of the Web resides in its capability of redirecting the information flow via hyperlinks, so it should appear natural that in order to evaluate the information content of a Web object, the Web structure has to be carefully analyzed". According to this assumption, many researchers investigated the role of links in IR. In particular, links have been used to (i) enhance document representation (104), (ii) improve document ranking by propagating document score (51), (iii) provide an indicator of popularity (29), and (iv) find hubs and authorities for a given topic (35). To our best knowledge, the task proposed in this section is the first study aimed at investigating the impact of semantically related links on CA.

It is worth noting that a key problem in this research field is how to measure the semantic relatedness of documents, see (31) for a survey. In this work, we just propose an experimental study on the impact of related links in CA without taking into account this problem.

As related links of the target webpage p, we consider: its *inlinks* (also called *backlinks*), i.e., pages that link to p; and its *outlinks* (also called *inbound* and *outbound links* depending on the corresponding domain), i.e., pages that are linked by p. Figure 6.2 gives a view on the considered related links, in which inlinks are represented by blue arrows and outlinks by red arrows. Two kinds of inlinks and outlinks may exist: those that link to an external domain (i.e., from A and to B in the Figure) and those that link to the same domain of the target webpage (i.e., from $T1$ and to $T2$ in the Figure). In this task, we consider only links belonging to a different domain. In other words we disregard inlinks that come from the same Web domain and inbounds (i.e, in the example in the Figure, we do not consider $T1$ and $T2$).

6.3.2 The proposed System

In order to study the role of related links in CA, we adopted the model depicted in Figure 6.3. Our model involves four modules: *Related Link Extractor*, *Text Summarizer*,

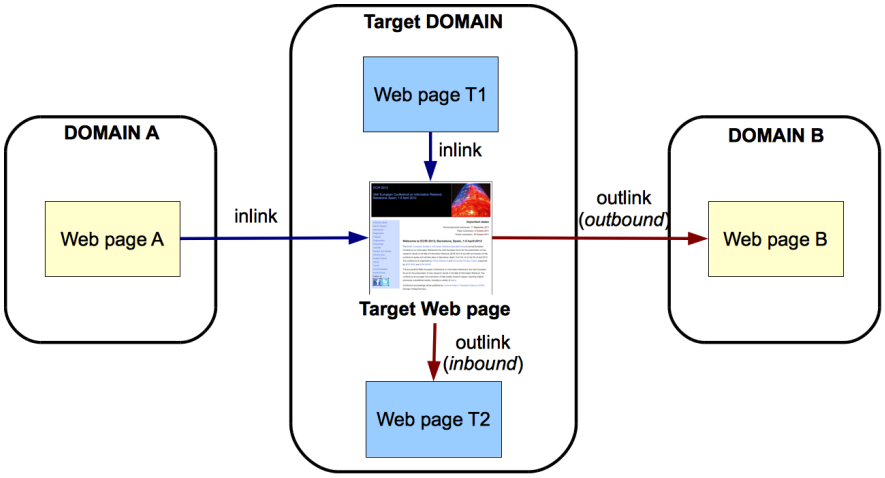

Figure 6.2: A graphical view of the adopted related links.

Classifier, and *Matcher*. The model in Figure 6.3 is compliant whit that in Section 3.1, in which only CF are considered in the matching phase. In particular, they coincide if the *related link extractor* is kept off.

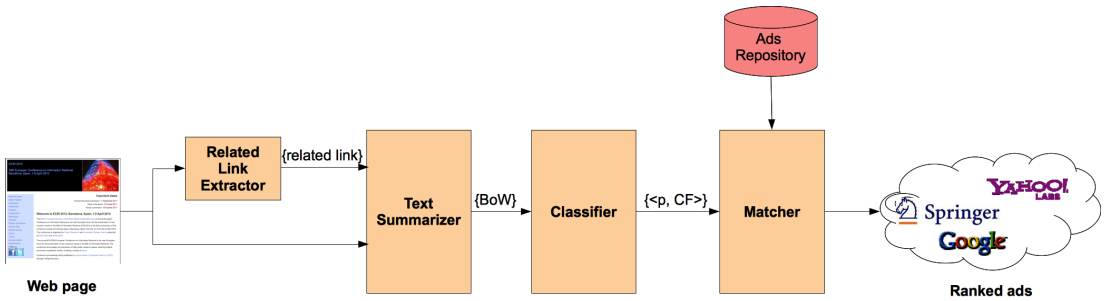

Figure 6.3: The model of the adopted approach.

Related Link Extractor

This module extracts the related links of a given webpage p. It collects: the first 10 inlinks of a given page by querying the Yahoo! search engine (http://www.yahoo.com)[1]; and the first 10 outlinks, if available, of a given page by parsing p looking for the anchor tag $< a >$[2].

[1]Further solutions may be adopted, including the use of existing tools, e.g., Page Inlink Analyzer (http://ericmiraglia.com/inlink/).

[2]To minimize the impact of "general-purpose" websites, we filter links such as Facebook, Twitter, Google+, eBay, and so on.

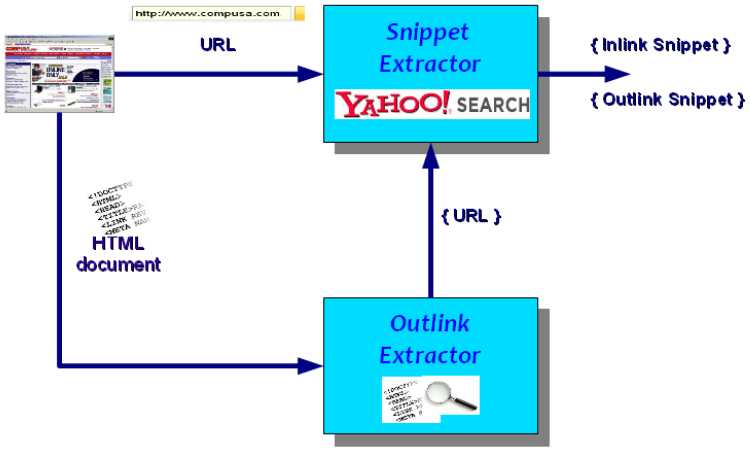

Figure 6.4: Related Link Extractor.

The input of the system is the webpage, including its URL, represented by its HTML code. The coded page is processed by the sub modules depicted in Figure 6.4

- ***Outlink Extractor*** Given the webpage code, the module parses it and extracts the set of outlinks. The URLs are sent to the Snippet Extractor.
- ***Snippet Extractor*** It is devoted to query the Yahoo! search engine in order to extract the snippets. It takes in input both the URL of the testing webpage and the set of URLs of the outlinks. The former is used to perform a special query to the Yahoo! search engine, in order to ask for the inlinks; from the resulting page the first 10 inlinks, filtered by skipping the links with the same domain, are selected and the associated snipped extracted. As for the latter, each URL are used to perform a conventional query to the search engine, the query being the entire URL; from the resulting page the first link suggested (usually is the actual webpage) is selected to extract the snippet. The output of the system is the set of snippet belonging to the selected inlinks or outlinks.

Text Summarizer

Instead of considering the whole page, we first extract the snippet of p, its inlinks and its outlinks by asking to Yahoo!. Therefore, the main purpose of this module is to

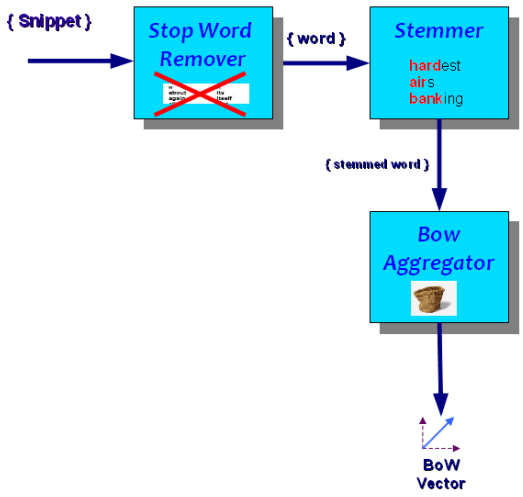

Figure 6.5: Text Summarizer.

process each snippet in order to remove stop-words and to stem each term through the Porter's algorithm (105). For each snippet, this module outputs a vector representation of the original text as BoW, each word being weighted by TFIDF.

The Figure 6.5 describes in details the module. It is built as the text summarizer adopted for the baseline system described in Chapter 3 without the modules devoted to parse, tokenize, and select the blocks of text.

- ***Stop word Remover.*** It takes the set of snippets and remove the stop words.
- ***Stemmer.*** For each term it stem it with the Porter's algorithm.
- ***BoW Aggregator.*** It outputs the bag of words in a vector-based representation.

Classifier

To infer the topics of each inlink, of each outlink, and of the original page, snippets are classified according to a given taxonomy. First, for each node of the taxonomy we merge all its documents into a single compound document. We then use it as a centroid for the Rocchio classifier (114) with only positive examples and no relevance feedback, as described in Section 5.1.2. Each centroid is defined as a sum of the TFIDF values of each term, normalized by the number of documents in the class. Snippet classification is based on the cosine of the angle between the snippet and the centroid of the class. The classifier outputs the snippet-category matrix, whose generic element w_{ij} reports the score given by the classifier for the category j to the snippet i.

Its main modules are sketched in Figure 6.6.

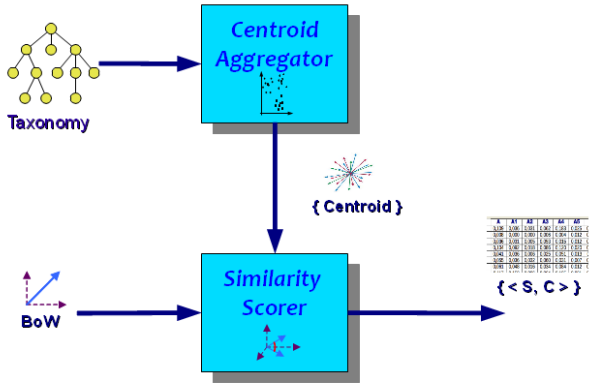

Figure 6.6: Classifier.

- ***Centroid Aggregator.*** The same module described for the Classifier of the baseline system (see Chapter 3).
- ***Similarity Scorer.*** It outputs the snippet-category matrix, each row being the selected snippet, and its cells being the scores associated to each category of the taxonomy.

Matcher

This module is devoted to suggest ads to the webpage according to the corresponding categories, i.e., the content. First, for each column of the page-category matrix, the matcher calculates the sum of the scores:

$$\sigma_i = \sum_{j=1}^{N} w_{ij} \tag{6.2}$$

where N is the total number of extracted pages, i.e., $N = 1 + N_i + N_o$, where N_i and N_o are the number of inlinks (fixed to 10) and the number of outlinks (which depends on the selected webpage, the maximum number is fixed to 10), respectively. Then, the *Matcher* selects k categories, i.e. those with the highest values of σ, k depending on

the agreement between publisher and advertiser. Finally, for each selected category, an ad is randomly extracted from the *Ads repository*[1].

6.3.3 Experimental Results

To assess if related links are effective for CA, we made experiments aimed at evaluate the contribution of the related links. Due to the twofolded nature of a related link (inlink or outlink), we also evaluate the individual contributions of the inlinks and outlinks.

For each experiment the evaluation is performed following two steps: first, the $\langle page,\ ad \rangle$ pairs are automatically judged by the adoption of the relevance score described in Chapter 3, and then the precision is computed considering as True Positive the relevant and somewhat relevant judgments. The Recreation dataset is adopted for the comparisons (see Chapter 3).

The Impact of Inlinks

The first step is to evaluate the contribution of the inlinks. To this end, we adopt the system described in Section 6.3.2 in which the Related Link Extractor relies only on the Snippet Extractor, and takes in input only the URL of the webpage. The output of this module is the set of inlink snippets. In order to evaluate the system, three experiments are performed by taking into account:

- the webpage alone (**P**), being compliant with the most part of state-of-the-art systems this is also the baseline;
- the set of inlinks alone (**I**);
- the webpage in conjunction with its inlinks (**P+I**).

For each system, the results in term of $\pi@k$, varying k, are depicted in Figure 6.7. Results show that the best performances are always obtained by ***P+I***. They also put into evidence that the main contribution is given by the introduction of inlinks. In particular, the introduction of inlinks with respect to the sole page (i.e., ***I*** vs. ***P***) leads to an increment of about 7.3%, whereas the introduction of the page with respect to the sole inlinks (i.e., ***P+I*** vs. ***I***) leads to an increment of about 0.6%. This means that the assumption that linked documents have related content is correct.

[1]We are considering to have already a repository in which the ads, i.e., company or service webpages, are classified according to the given taxonomy.

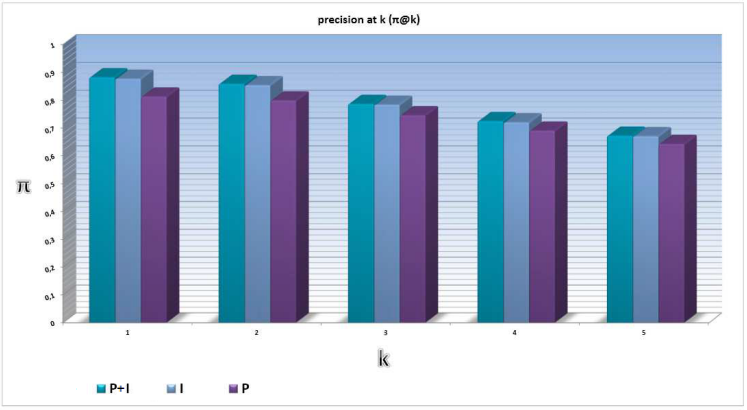

Figure 6.7: *Inlinks:* precision at k.

To show the behavior of each system with respect to each category of the taxonomy, Figure 6.8 depicts the π@1 calculated category per category. Also these results show that ***P+I*** and ***I*** perform always better that the baseline system. The best performances are obtained for the *Travel* sub-tree (T1,...T5), except for for the *Travel* sub-root category (T) in which ***P*** overwhelms both ***I*** and ***P+I***. Let us also note that all the systems showed the worst performances for the *Aviation* and *Collecting* sub-root categories (A and C).

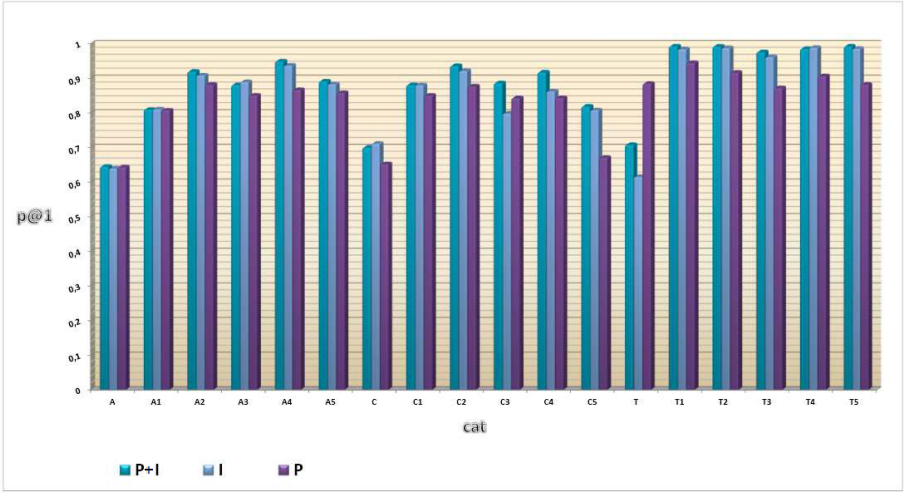

Figure 6.8: Inlinks: precision at 1 for each category.

The Impact of Outlinks

We adopt the system described in Section 6.3.2 in which the Snippet Extractor provides the snippets of the outlinks together with the single snippet of the webpage. Three performative experiments are performed by taking into account:

- the webpage alone (**P**), being compliant with the most part of state-of-the-art systems this is also the baseline;
- the set of outlinks alone (**O**);
- the webpage in conjunction with its outlinks (**P+O**).

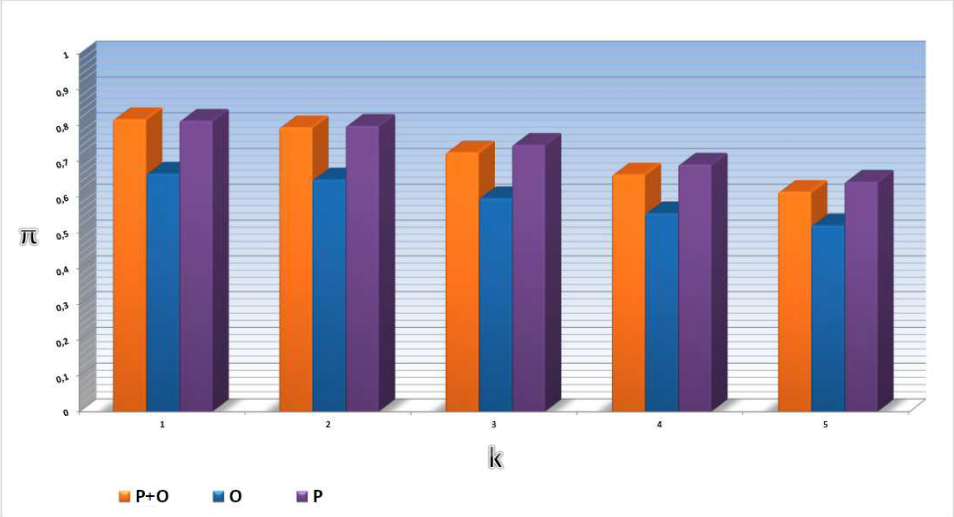

Figure 6.9: *Outlinks:* precision at k.

The results in term of $\pi@k$, varying k, are depicted in Figure 6.9. Results show that the best performances are obtained by ***P***. So that, the adoption of outlinks leads to a remarkable decrease of the performances. The differences between the adoption of inlinks and the adoption of outlinks could be due to the different amount of examined links. In fact, for a given webpage, even if it is easy to find at least 10 inlinks, it is more difficult that a webpage contains at least the same number of outlinks. We analyzed the dataset in order to compute the number of outlinks per page, and we found that a webpage contains an average number of 3.6 inlinks.

Figure 6.10 depicts the $\pi@1$ calculated category per category. Also these results show a different behavior of the *Travel* sub-root category (T), and the best performances

obtained for the *Travel* sub-tree (T1,...T5).

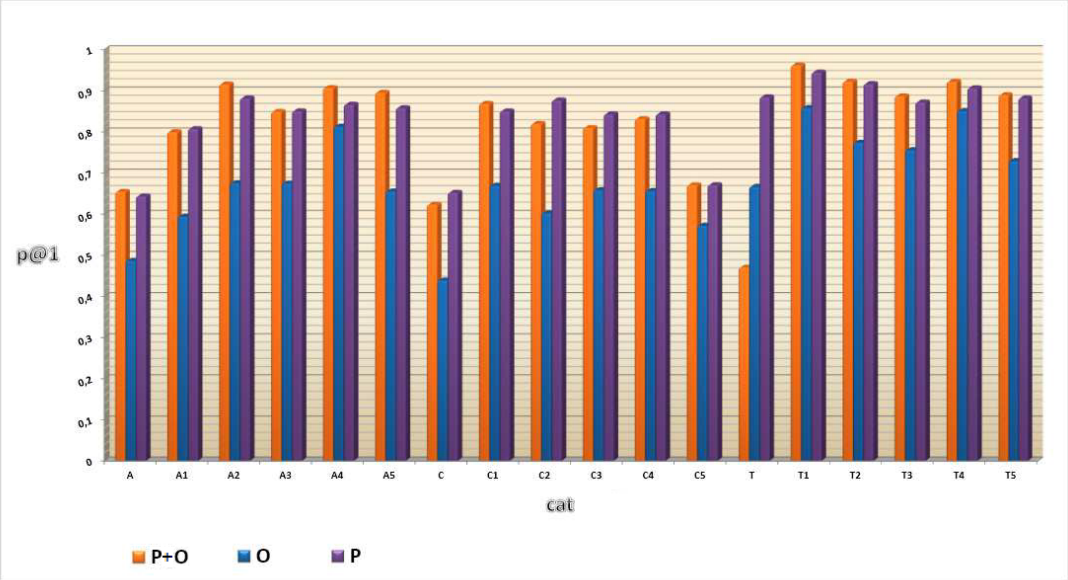

Figure 6.10: Outlinks: precision at 1 for each category.

The Impact of related Links

We performed the experiment aimed at evaluate the conjuncted adoption of inlinks and outlinks. The adopted configurations take into account:

- the webpage alone (**P**);
- the set of related links (inlinks and outlinks) alone (**RL**);
- the webpage in conjunction with its related links (**RL**).

The results in term of $\pi@k$, varying k, are depicted in Figure 6.11. Results show that the best performances are obtained by ***P***. So that, the adoption of outlinks leads to a remarkable decrease of the performances. Results show that the best performances are always obtained by ***P+RL***. The introduction of related links with respect to the sole page (i.e., ***RL*** vs. ***P***) leads to an increment of about 9.2%, whereas the introduction of the page with respect to the sole inlinks (i.e., ***P+RL*** vs. ***I***) leads to an increment of about 1.4%.

Figure 6.12 depicts the $\pi@1$ calculated category per category. These results are compliant with those showed for ***P+I*** and ***P+O***. They show a different behavior of

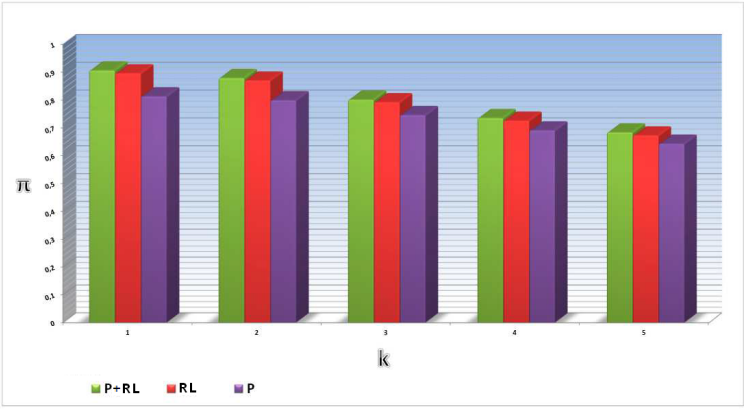

Figure 6.11: The *Related links: precision at k*.

the *Travel* sub-root category (T), and the best performances obtained for the *Travel* sub-tree (T1,...T5).

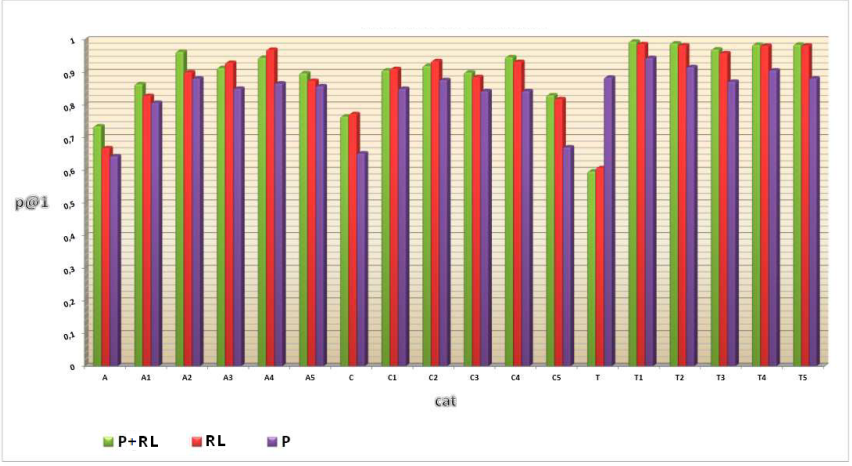

Figure 6.12: Related links: precision at 1 for each category.

To better assess that, even if outlinks alone negatively affect the system, the adoption of both inlinks and outlinks have a positive impact we report in the Figure 6.13 the precisions of each system. In the figure it s more clear how the adoption of related links increases the performances.

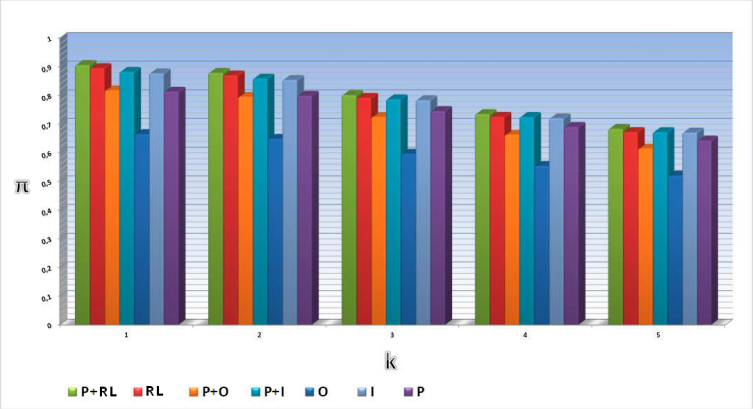

Figure 6.13: Precision at k.

The Impact of Somewhat Relevant

To verify how performances are affected by the choice of selecting as TP both ads scored as 1 or 2, we also studied the impact of somewhat relevant $\langle p,\ a\rangle$ pairs with respect to relevant ones.

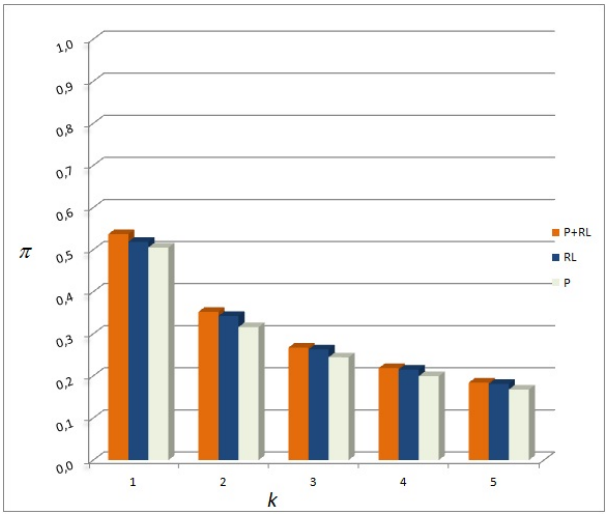

Figure 6.14: The precision calculated considering as TP only pairs scored by 1.

Figure 6.14 depicts the *precision at k* calculated considering as TP only $\langle p,\ a\rangle$ pairs scored by 1, i.e., somewhat relevant pairs are considered as FP. The Figure shows that exploiting related links slightly improves the precision and that the best results are obtained considering both the page and the related links.

Summarizing, results show that the adoption of the information provided by related

links improves the performance of the proposed system. In particular, the *precision at* k of ***P+RL*** is always better than that obtained by ***P*** and by ***RL*** along the values of k and no matter whatever TP contains somewhat relevant pairs or not.

An Example

To better clarify the adopted approach, let us consider as target page the home page of the portal Vans Air Force[1], a site catering to the lifestyle of building and flying Van's Aircraft RV kit planes. The webpage is categorized as *Aircraft* in DMoz. A fragment of the home page is depicted in Figure 6.15.

Figure 6.15: An example of target page.

First, the CA system queries Yahoo! asking for the first 10 inlinks and extracts the outlinks. Subsequently, the corresponding snippets are pre-processed and classified, the output being a page-category matrix. Figure 6.16 shows the page-category matrix for the given page together with the sum of the scores for each category. Upon the choice of suggesting 3 ads, the three most significant categories are *Aviation* (A), *Organizations* (A4), and *Aircraft* (A1). For each of these categories, the system suggests an ad randomly extracted from the *Ads repository*: an online shop for jets and air kits (http://www.rbckits.com/shop/), for the category *Aviation*; an online shop for supplier of radio controlled helicopters, airplanes and cars (http://www.toyrc.com/), for

[1] http://www.vansairforce.net/

the *Organizations* category; and a full service company specializing in acquiring, reactivating, and ferrying transport category aircraft (http://www.globalaircraftllc.com/), for the *Aircraft* category.

	A	A1	A2	A3	A4	A5	C	C1	C2	C3	C4	C5	T	T1	T2	T3	T4	T5
Van's Aircraft RV Builder	0,109	0,036	0,031	0,062	0,183	0,025	0,032	0,027	0,014	0,010	0,025	0,012	0,051	0,017	0,024	0,030	0,080	0,031
Van's RV Italia	0,008	0,000	0,000	0,006	0,004	0,012	0,005	0,005	0,000	0,000	0,008	0,000	0,017	0,006	0,002	0,000	0,038	0,019
Karmy Flying Adventures' Video	0,036	0,031	0,005	0,053	0,015	0,012	0,012	0,010	0,003	0,005	0,004	0,010	0,024	0,015	0,033	0,012	0,015	0,002
Tony Phillips RV9 Build Log	0,104	0,092	0,018	0,086	0,120	0,020	0,041	0,025	0,024	0,011	0,026	0,031	0,055	0,015	0,054	0,029	0,034	0,050
Aircraft Tools	0,041	0,036	0,006	0,025	0,051	0,013	0,011	0,002	0,018	0,002	0,002	0,012	0,033	0,010	0,005	0,047	0,055	0,000
Vans RV8	0,055	0,036	0,032	0,060	0,031	0,007	0,014	0,010	0,010	0,006	0,008	0,002	0,021	0,001	0,010	0,007	0,041	0,021
Doug Reeves' Van's Air Force Web Site	0,061	0,048	0,016	0,034	0,084	0,012	0,004	0,005	0,002	0,004	0,000	0,000	0,022	0,002	0,019	0,019	0,032	0,007
Flying - Philip Greenspun's home page	0,117	0,153	0,006	0,064	0,135	0,031	0,009	0,005	0,008	0,005	0,002	0,006	0,035	0,003	0,011	0,079	0,024	0,001
Bruce Swayze's RV-7A Project	0,007	0,007	0,002	0,008	0,004	0,002	0,008	0,004	0,006	0,003	0,001	0,016	0,008	0,002	0,009	0,009	0,005	0,082
The RV Aircraft Journal	0,033	0,042	0,015	0,026	0,012	0,009	0,024	0,018	0,016	0,006	0,012	0,019	0,032	0,019	0,020	0,017	0,044	0,009
SDS EM-4: Aircraft	0,092	0,049	0,050	0,093	0,062	0,021	0,074	0,043	0,053	0,028	0,038	0,050	0,055	0,012	0,022	0,034	0,062	0,063
Dynon Avionics	0,018	0,019	0,005	0,013	0,006	0,014	0,012	0,003	0,001	0,012	0,009	0,004	0,093	0,030	0,053	0,062	0,027	0,135
N617AR.COM	0,048	0,062	0,001	0,066	0,034	0,004	0,008	0,007	0,004	0,002	0,005	0,006	0,014	0,007	0,018	0,005	0,015	0,000
VFR Terminal Area Raster Aeronautical Charts	0,049	0,049	0,011	0,045	0,040	0,013	0,069	0,036	0,050	0,017	0,039	0,062	0,045	0,025	0,039	0,017	0,017	0,048
XCOM 760 VHF Aircraft Transceiver for Sport A	0,023	0,007	0,002	0,033	0,027	0,005	0,013	0,008	0,007	0,008	0,004	0,009	0,012	0,002	0,012	0,009	0,001	0,015
Bonaco Inc.	0,024	0,016	0,010	0,029	0,012	0,008	0,022	0,012	0,024	0,004	0,014	0,011	0,015	0,011	0,010	0,008	0,010	0,009
NEW GRT Avionics Autopilot	0,049	0,031	0,012	0,048	0,044	0,020	0,056	0,031	0,050	0,010	0,035	0,041	0,051	0,038	0,044	0,006	0,038	0,038
SteinAir, Inc. Custom Panels, Builders ...	0,004	0,000	0,000	0,013	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
FatBoyz Aviation	0,158	0,121	0,030	0,155	0,173	0,031	0,016	0,010	0,015	0,005	0,006	0,012	0,043	0,003	0,088	0,007	0,038	0,013
Lycoming Fuel Injection and Ignition	0,030	0,032	0,006	0,024	0,020	0,014	0,026	0,010	0,021	0,003	0,021	0,024	0,030	0,004	0,007	0,008	0,024	0,063
Andair's Web Site	0,023	0,011	0,009	0,019	0,025	0,008	0,034	0,016	0,020	0,008	0,025	0,030	0,043	0,023	0,025	0,013	0,024	0,058
SCORES	**1,091**	**0,879**	0,267	0,960	**1,081**	0,281	0,492	0,288	0,347	0,151	0,283	0,357	0,698	0,246	0,505	0,418	0,626	0,664

Figure 6.16: The page-category matrix.

6.4 A Recommender System based on a Generic Contextual Advertising Approach

As discussed in Section 6.2 RS and CA could be seen as a unifying view. On the one hand, CA systems are devoted to suggest suitable advertising to users while surfing the web (43). On the other hand, RS are devoted to suggest interesting items to users, as showed in Section 6.1.2. In this perspective, CA is a type of Web recommendation, which, given the URL of a webpage, aims to embed into the page the most relevant textual ads available.

This new setting permits us to investigate the possibility of devising a RS *a la mode of* CA (3; 4). To our best knowledge, this is the first attempt to exploit CA techniques to devise and implement a RS.

Taking into account the baseline system for CA described in Section 3.1, we could see the similarities between it and a generic context-based RS. Our approach in building a RS involves two steps: user profiling and recommendation. For each user of the system, a profile is generated from a set of documents rated as relevant by the selected user, i.e., the user history. New documents can then be proposed to the user and added to her/his user history if they match the corresponding profile.

User Profiling

Figure 6.17 sketches the architecture of the proposed system, composed by four main modules: statistical document analyzer, semantic analyzer, semantic net handler, and profiler (5).

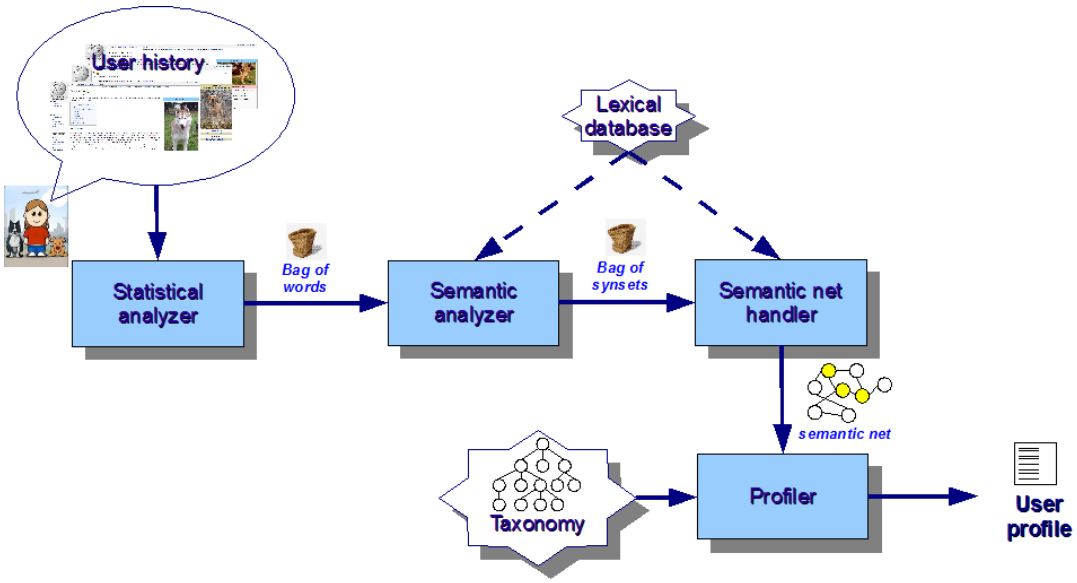

Figure 6.17: The user profiler at a glance.

Statistical document analyzer. While analyzing documents rated as relevant by the user, this module is devoted to create the bag of words (BoW). The BoW collects all terms contained in the input documents, suitably weighted. The statistical document analyzer removes from the BoW all non-informative words such as prepositions, conjunctions, pronouns, and very common verbs using a stop-word list. Subsequently, it calculates the weight of each term adopting the TFIDF measure. The statistical document analyzer calculates an overall TFDIF considering all documents in D, being D the user history. Furthermore, the weights resulting from TFIDF undergo a cosine normalization. To reduce the dimensionality of the space, only the first N terms of the BoW are retained. The optimal value of N must be calculated experimentally; in this work, good values have been found in the range 60-90 (due to the fact that –here– textual descriptions are often short). Hereinafter, the set of terms stored in the BoW will be called *features*. This module corresponds to the text summarizer adopted in the generic CA solution described in Section 3.1.

Semantic analyzer. This module creates the bag of synsets (BoS), which collects all synsets related to the selected features. To this end, the semantic document analyzer

queries the online lexical database WordNet (99). In particular, for each feature, WordNet provides the corresponding synsets. After synset extraction, the semantic document analyzer assigns to each synset a weight according to the TFIDF of all related terms.

This module corresponds to a text summarizer based on semantic information. In fact, a semantic approach can be also adopted in CA to improve the performances of the text summarization task.

Semantic net handler. This module aims to (i) build the semantic net from the BoS and (ii) extract its most relevant nodes. First, the semantic net is built in form of a graph, where nodes are the synsets belonging to the BoS, and edges are semantic relations between synsets. Four kinds of semantic relations are taken into account: *hyponymy* (sub-name) and its inverse, i.e., *hyperonymy* (super-name); *meronymy* (has-part) and its inverse, i.e., *holonymy* (member-of).

The semantic net handler is also in charge of pruning the network by dropping not relevant nodes, identified according to their weight and to the number of connections with other nodes.

Profiler. This module is devoted to extract the user profile. To this end, it exploits the WordNet Domains Hierarchy (WNDH) (86) and associates the proper category to each selected node[1].

Considering the selected nodes, together with their weights, the profiler is able to identify the real interests of a user in terms of WNDH. In particular, a user profile is represented by a set of pairs $\langle c_k, w_k \rangle$, where c_k is a WNDH category and w_k is the corresponding weight in $[0, 1]$. The semantic net handler and the profiler correspond to the classifier adopted in the generic CA solution previously described.

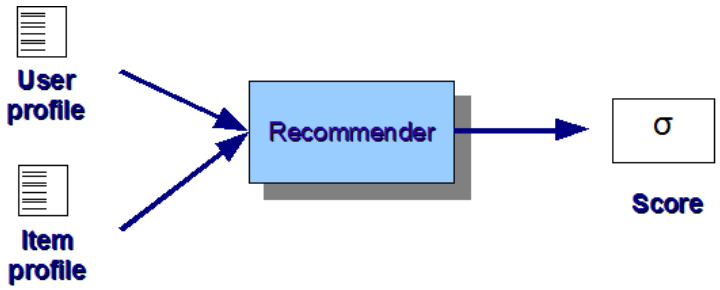

Figure 6.18: The recommender system at a glance.

[1]The actual taxonomy can be selected according to the corresponding application field.

Recommendation

Once the user profile has been generated, the system can rank a new item I to determine whether it could be of interest for a specific user (see Figure 6.18). This can be done by measuring the distance between the vector-based representation of the item I, $\vec{V}(I)$, and the user profile vector U, $\vec{V}(U)$. In particular, the textual information of a given item is processed in a way similar to that of profile extraction: a set of WDNH categories with the corresponding relevance ratio are computed for the item, and the cosine distance between U and I is evaluated. In formula:

$$sim(I,U) = \frac{\vec{V}(I) \cdot \vec{V}(U)}{|\vec{V}(I)\|\vec{V}(U)|} \tag{6.3}$$

where, the numerator represents the dot product of the vectors $\vec{V}(I)$ and $\vec{V}(U)$, and the denominator is the product of their Euclidean lengths. Items obtaining a score greater than 0.5 are proposed to the user. It is easy to note that the recommender corresponds to the matcher adopted in the generic CA solution previously described.

6.4.1 Experimental Results

To assess whether our approach can be adopted to profile users in terms of the categories they are interested in, we first resorted to the approach described in (2) to build a dataset in which documents are classified according to WNDH categories. Selecting documents from such dataset allowed us to automatically create user histories. Several experiments have been performed, averaging on the number n of categories (ranging from 1 to n) and keeping fixed (at a rate $1/n$) the amount of documents belonging to each category. Results, calculated for different numbers N of features (from 10 to 100), show that the best result is obtained for N=80.

Subsequently, we performed also a preliminary study about the impact of changing the number of documents associated to a given category in a user history. As a starting point for studying such phenomenon (say category imbalance), only two categories have been considered. In particular, for each combination of two categories (say, A and B), experiments have been performed starting from 5% of documents belonging to A and 95% to B, incrementing such percentage up to 95% for A and vice versa for B (with a delta of 5%). Comparing such percentages with those given as output by the system, we calculated the MSE. Experimental results point out that the filtering activity of the Statistical Document Analyzer is more effective when the user history is composed by a large amount of documents of a specific class. In our view, this is due to the fact

that a strong bias on a given category facilitates the system in the task of identifying it. Moreover, the fact that in this case the remaining category has a lower range of variation has also a positive impact. Let us also note that this result is obtained by averaging all pairwise combinations of categories, without taking into account that some categories may be correlated (e.g., *Medicine* and *Biology*).

The experimental setting is concerned with assessing the mean average precision of: i) the proposed RS say *CB*, ii) an item-based system, say *IB*, compliant with the state-of-the-art algorithm proposed in (119), and iii) an hybrid system, say *H*, that embeds both. To this end, we devised a web application, i.e., a photo recommender service, in which a set of 700 photos [1] have been presented to 26 users, who were asked to express a rate from 1 to 5 for each photo according to their appreciation (1 means "not interesting at all", 5 means "very interesting").

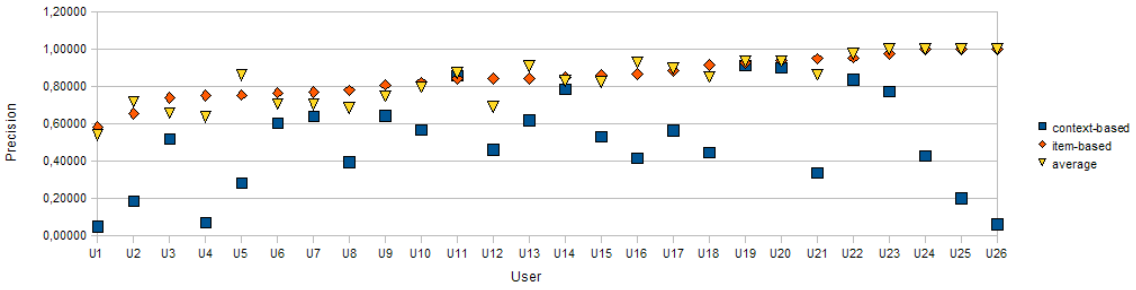

Figure 6.19: Experimental results in terms of precision, recall, and accuracy.

The overall activity consists of 4 steps. The first step is devoted to collect preliminary information about users: to this end each user is requested to rate 100 photos randomly selected from the 700 available. Once the first step is completed, the information is sent to *CB* and *IB*. Such information is used by the former to generate user profiles and by the latter to generate the initial user-item matrix. In the subsequent step *CB* and *IB* select the 20 photos with the highest rating and the web application displays them in random order. At this point, each user is required to rate the displayed photos. The step is considered completed only when all users have rated the whole set of 40 photos. Two additional steps, with the same characteristics of the one described above, are performed with the goal of incrementally estimating the precision of each recommender in isolation from the other, as well as the precision of *H*. It is worth

[1] The photos were downloaded from Flickr, among those fulfilling Creative Commons License and having a title and a description which had to be related to some of the contexts included in WNDH.

pointing out that the score of H is calculated by averaging the score of CB and IB according to a weighting schema that accounts for the corresponding precisions (each weight being in fact a normalized precision, updated at each step).

Figure 6.19 shows the results in terms of precision for each user. The hybrid approach produces interesting results for some users, but not for all of them. The main reason is due to the constraining context, which is not particularly suitable for a context-based algorithm, since the textual descriptions of photos are scarce and often ambiguous. Moreover, most of the users probably rated more the beauty of the photos rather than the subjects.

Chapter 7

A Preliminary Study on Geo Targeting for Online Advertising

In the Internet marking, geo targeting is the task of determining the geolocation of a website visitor in order to provide different content according to him/his location, such as, country, state, or zip code. A common usage of geo targeting is found in online advertising, in order to suggest localized ads. In this chapter we investigate the background and propose a preliminary study on the users' behaviors in the field of mobile services.cation, such as, country, state, or zip code. A common usage of geo targeting is found in online advertising, in order to suggest localized ads. In this chapter we investigate the background and propose a preliminary study on the users' behaviors in the field of mobile services.

7.1 Geo Targeting: Background

Content localization nowadays has an important role in the World Wide Web. For example, someone who types the word "cars" into a search engine should be interested in only receiving results on cars in his local area. Another application is when a Web portal like Yahoo! dynamically renders content in the user language, potentially showing local news, traffic, weather and other data that is specifically pertinent to him.

Several studies have been made recently to infer geographic information from the textual content. The task of finding a geographical focus of a webpage was first proposed by Ding et al. (46). Their approach was two-fold: finding locations of webpages with hyper-links to the analyzed page and detection (and disambiguation) of toponyms in its content. Works by Amitay et al. (8) and Zong et al. (142) relied on propagating the

confidence weights of found toponyms up to the root of a taxonomy to find the most probable common ascendants (e.g., finding a country for several cities mentioned).

Working with blog data, Mei et al. (93) presented methods to find latent semantic topics and their distribution over locations (states or countries), whereas Wang et al. (131) propose a Location-Aware Topic Model based on Latent Dirichlet Allocation.

Focusing on the Web queries, they are more similar to tags than blogs, in the sense that queries consist on two or three content terms representing much larger concepts. Backstrom et al. (19) proposed a method to measure the geo-specificity of a query, using the level of dispersion around the location of the query's highest frequency. With a similar goal in mind, Zhuang et al. (140) calculated the inverse correlation of a query click distribution over locations with their populations. Vadrevu et al. (130) use the probability of co-occurrence of a query term with place names from each region to determine queries that might be related to a given region.

7.2 Geo Targeting in Online Advertising

The main goal of advertising is to increase revenue of the advertisers by improving the user's experience. To improve the effectiveness of the advertising process is useful to study the users' behavior by relating to the characteristic of them. Understanding the geographic context of a generic user could be useful for providing several services to him. This task should be exploited by advertisers for suggesting local (or hyper-local) ads.

As a matter of fact, the evolution of the technology is going toward a continuous growth of Web and mobile services use by common people. Nowadays the use of Web services for mobile applications is the easiest way in the search of every kind of needs, including the hyper-local needs (e.g., a search for the closest market).

The ability to sell and place ads based on city level geography has, in a powerful way, brought more advertisers to the online medium. If someone owned a local restaurant in Barcelona he would not really benefit from all the generalized Internet traffic coming from Madrid or Bilbao, for example. He would want to buy just city level impressions online.

There is, therefore, a pressing need for several novel geo-localized services and tools that can be useful for increasing the revenue of advertisers and publishers.

7.3 A Preliminary Study on Mobile Queries

The task related to the research activity is focused on the preliminary studies of the interaction between the user and the suggested items in a mobile search engine. The currently opened analyses are related to how and in what measure a query could be considered as "local" query, in order to suggest, for a specific local mobile search, an ad related to the user's intent.

To this end, the first step has been the analyses of the distributions of distances between the user stated profile location and the location of him/his at the moment of the requested query.

7.3.1 Yahoo! Placemaker

We adopt the Web service Yahoo! Placemaker [1] to extract geographic information. Yahoo! Placemaler is a service able to provide geographic information from structured and unstructured textual elements, such as webpages, RSS feeds, blogs, or news articles. Placemaker is a suitable instrument to infer geographically relevant content that is directly not discoverable. It identifies and disambiguates places mentioned in the text, providing unique identifiers (WOEIDs) for each place. In simple terms, Placemaker finds what places are referenced in a text segment.

The Placemaker API is accessed via HTTP POST , with some parameters to specify, and it outputs an XML document containing all the geographic information. For example, a possible request to the API is the following:

documentContent=Sunnyvale+CA
documentType=text/plain
appid=my_appid

and the service produces the following result:

```
<?xml version="1.0" encoding="utf-8"?>
  <contentlocation
    xmlns:yahoo="http://www.yahooapis.com/v1/base.rng"
    xmlns:xml="http://www.w3.org/XML/1998/namespace"
    xmlns="http://wherein.yahooapis.com/v1/schema"
    xml:lang="en">
      <processingTime>0.001432</processingTime>
        <version> build 090409</version>
```

[1] http://developer.yahoo.com/geo/placemaker/

```
<documentLength>12</documentLength>
  <document>
    <administrativeScope>
      <woeId>2502265</woeId>
      <type>Town</type>
      <name><![CDATA[Sunnyvale, CA, US]]></name>
      <centroid>
        <latitude>37.3716</latitude>
        <longitude>-122.038</longitude>
      </centroid>
```

...

7.3.2 Experiments

We are interested in studying the correlation between the user location and the location of the clicked service suggested by the mobile search engine. In particular, the interest is in the URLs suggested by the search engine. For each resulting link is in fact possible estimate the geographic location of the service suggested by the search engine by analyzing the Web link. The estimation is obtained by the adoption of the geo service Yahoo! Placemaker. In particular, from the Placemaker API two locations are extracted for each URL: the GeoScope location and the smallest place location returned by the service. So, two distances are computed for each URL: the distance between the user's location and the GeoScope location (d_G) and the distance between the user's location and the smallest place location (d_S).

Data Gathering

From the Yahoo! Grid the data related to mobile queries have been selected. First, the clicked queries of the month of October 2011 have been filtered. Only the queries with the following parameters are considered:

- 4 or more characters per query;
- Coordinates not nulls;
- Number of distinct clicks greater than 10 (clicks by different users).

A small test set has been extracted for the preliminary studies (about 5000 queries). For each query the distances d_G and d_L are computed (one for each click). The total clicks have been about 78000.

The current experiments consist on the evaluation of the distribution of clicks by the distances, in terms of total distribution, by the single click, or the relative distribution, considering the relative distances for each query.

The dataset is built as a set of queries; each query has a set of pairs $< d_G, d_S >$, relative to each click. For both geoscope and smallest distance, the centroid is taken as geographic location of the place.

Global Analysis

Taking into account each click without grouping by queries, the distribution of distances is computed. The graph in Figure 7.1 shows the number of clicked URLs for a given distance. The step of x axes is 1 Km; for instance, the first step (1 km) means that for the range of 0-1 Km the number of clicks with a distance d_s included in that range is reported.

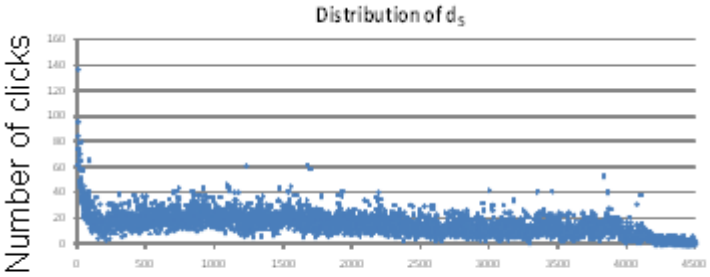

Figure 7.1: *Distribution of* d_S

Even if it seems a uniform distribution in the middle range (slightly decreasing), if we focus on the head of the graph we could see an initial peak, decreasing faster. By sketching the graph for the first 100 km, we could see that most clicked queries are concentrated in the first area of 1 km radius, and then decreases faster, as depicted in Figure 7.2

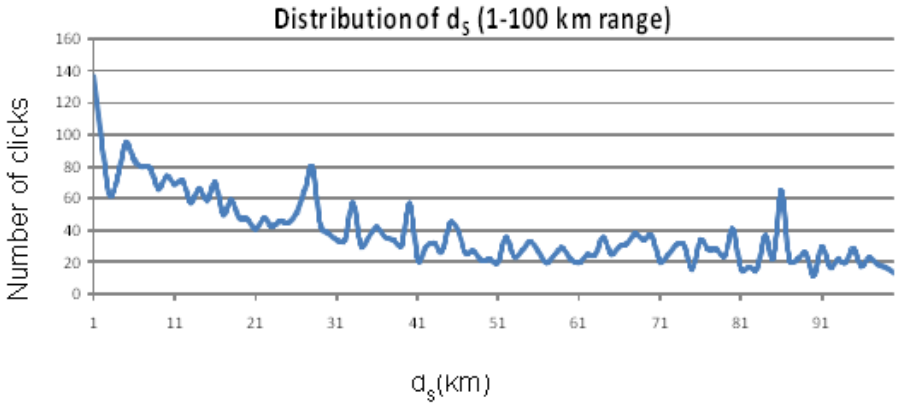

Figure 7.2: *Distribution of* d_S for the area of 100 Km

The same behavior could be seen if we observe the Geoscope distance (d_G) for the first 100 Km, as sketched in Figure 7.3

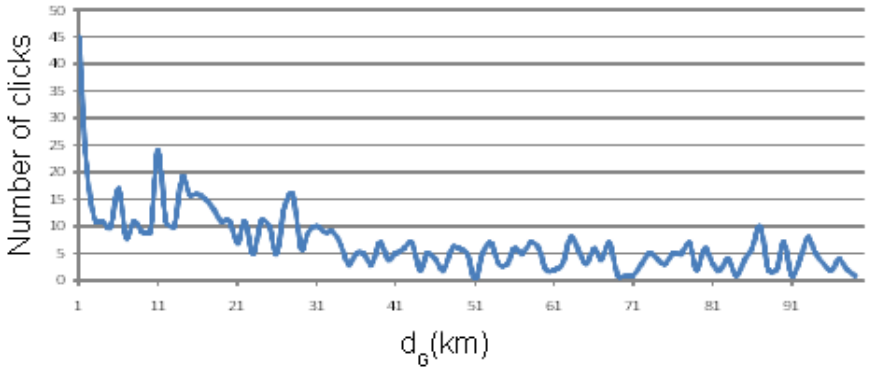

Figure 7.3: *Distribution of d_G* for the area of 100 Km

Query analysis

For each query we could analyze the behavior of clicks related to the distances. For each query, if we take the minimum and the maximum distance, and we divide this range into n equidistant intervals, we could report the associated interval for each click. We computed that for the whole set of queries, and we choose n = 100; the graphs depicted in Figure 7.4 and Figure 7.5 show respectively the behavior for d_S and d_G.

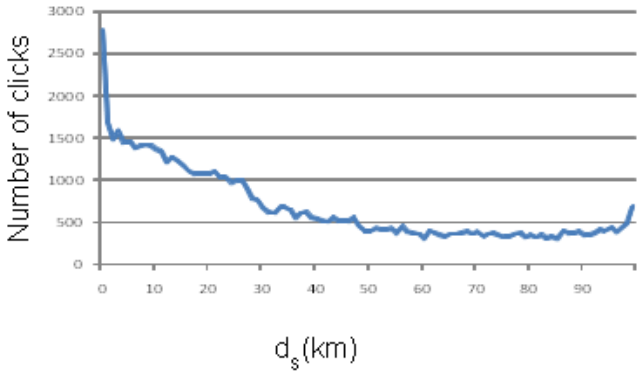

Figure 7.4: *Relative distribution of d_S*

From the graph in Figure 7.4 we could say that the clicked links mostly belongs to an area with a radius of 1% of the maximum area suggested for each query, if we take into account the smallest places of the resulting clicked link.

An interesting behavior is the one corresponding to the analysis of the geodistances (Figure 7.5), that shows an increasing of queries localized between 3 and 14% of the maximum radius

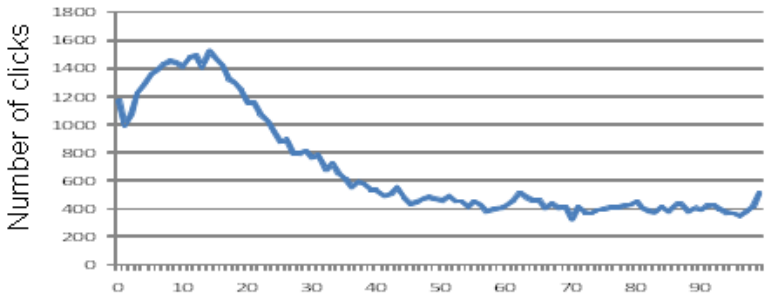

Figure 7.5: *Relative distribution of d_G*

for each query. This could it due to the fact that the Geoscope location is assigned for each link could be of any place type, and often could be related to a region, land, geographic zone, whereas the smallest place extracted has an administrative scope (from the district of a city to a country), and could be more "localized".

Let us note that both graphs show a sort of linear behaviors, that is index of a correlation between clicked documents and location of the user. From the graphs sketched in Figure 7.2 and Figure 7.3 we could say that users are often more interested in services with a local intent, and it is possible chracterize a subset of users with that intents. This aspects could be more clear if we consider the total views of the suggested links, and analyzing the distributions. This is currently under investigation.

7. A PRELIMINARY STUDY ON GEO TARGETING FOR ONLINE ADVERTISING

Chapter 8

Contextual Advertising on Multimodal Aggregation

Modern broadcasters are facing an unprecedented technological revolution from traditional dedicated equipment to commodity hardware and software components, and from yesterday-one-to-many delivery paradigms to nowadays-Internet-based interactive platforms. Thus, information engineering and integration plays a vital role in optimizing costs and quality of the provided services, and in reducing the "time to market" of data.

Nowadays, Web is characterized by a growing availability of multimedia data together with a strong need for integrating different media and modalities of interaction. Hence, the main goal is to bring into the Web data thought and produced for different media, such as TV or radio content.

In this challenging scenario, this chapter focuses on multimodal news aggregation, retrieval, and fruition. Multimodality is intended as the capability of processing, gathering, manipulating, and organizing data from multiple media (e.g., television, radio, the Internet) and made of different modalities such as audio, speech, text, image, and video. In our work, news come from two inputs: (i) automatically extracted, chaptered, and transcribed TV news, and (ii) RSS feeds from online newspapers and press agencies.

In particular, we tackle two main issues: to extract relevant keywords to news and news aggregations, and to automatically associate suitable advertisements to aggregated data. To achieve the first goal, we propose a solution based on the adoption of extraction-based text summarization techniques; whereas to achieve the second goal, we developed a CA system that works on multimodal aggregated data. To assess the proposed solutions, we performed experiments on Italian news aggregations.

G.Armano, **A.Giuliani**, A. Montagnuolo, A.Messina, and E.Vargiu. *Experimenting Text Summarization on Multimodal Aggregation*, DART'11: 5th International Workshop on New Challenges in Distributed Information Filtering and Retrieval, 2011.

G.Armano, **A.Giuliani**, A. Montagnuolo, A.Messina, and E.Vargiu. *Content-based Keywords Extraction and Automatic Advertisement Associations to Multimodal News Aggregations*. "New Challenges in Distributed Information Filtering and Retrieval", Studies in Computational Intelligence series, Springer-Verlag, C. Lai. G. Semeraro and E. Vargiu (in press), 2011.

8.1 Background

8.1.1 Information Fusion and Heterogeneous Data Clustering

Information (or data) fusion can be defined as the set of methods that combine data from multiple sources and use the obtained information to discover additional knowledge, potentially not discoverable by the analysis of the individual sources.

First attempts to organize a theory have been done in (106), in which the author proposes a cross-document structure theory and a taxonomy of cross-document relationships. Recently, some proposals have been made to provide a unifying view. The work in (69) classifies information fusion systems in terms of the underlying theory and formal languages. Moreover, in (87), the author describes a method (Finite Set Statistics) which unifies most of the research on information fusion under a Bayesian paradigm.

Many information fusion approaches currently exist in many areas of research, e.g., multisensor information fusion, notably related to military and security applications, and multimedia information fusion. In the latter branch, the closest to the present research, the work in (136) analyses best practices for selection and optimization of multimodal features for semantic information extraction from multimedia data. More recent relevant works are (103) and (76). In (103), the authors present a self-organizing network model for the fusion of multimedia information. In (76), the authors implement and evaluate a fusion platform implementing the framework within a recommendation system for smart television in which TV programme descriptions coming from different sources of information are fused.

Heterogeneous data clustering is the usage of techniques and methods to aggregate data objects that are different in nature, for example video clips and textual documents. A type of heterogeneous data clustering is co-clustering, which allows simultaneous clustering of the rows and columns of a matrix. Given a set of m rows in n columns, a co-clustering algorithm generates co-clusters, i.e., a subset of rows which exhibit similar behavior across a subset of columns, or vice-versa. One of the first methods conceived to solve the co-clustering of documents using word sets as features is represented by (80), where RSS items are aggregated according to a taxonomy of topics. More challenging approaches are those employing both cross-modal information channels, such as radio, TV, the Internet, and multimedia data (44; 137).

8.1.2 Multimedia Semantics

Recently, several research activities have attempted to provide the state-of-the-art of content-analysis-based extraction of multimedia semantics, with the stated intention to provide a unified perspective to the field (45; 57; 79; 125). Mostly, these works succeed in giving a complete and updated panorama of the existing techniques based on content analysis for multimedia knowledge representation. In our opinion, the work done so far has only partially achieved the objective of giving a deep understanding of problems related to multimedia semantics. This statement comes from the observation that only a very few research solutions and tools end up to be useful for practical purposes in the media industry. In our opinion, this is due to a significant lack of precision in the definition of relevant problems, which led to huge research efforts, but only seldom in directions exploitable by the media industry (e.g., broadcasters, publishers, producers) in a straightforward way. Emergent technologies like Omni-Directional Video (59) enhance the urgency of a high-level re-elaboration of the discipline.

Modern research efforts in Multimedia Information Retrieval (MIR) have been recently summarized by Lew et al. (79). One of the key issues pointed out by the authors is the lack of a common and accepted test set for researchers conducting experiments in the field of MIR. The somewhat central claim of Lew et al. is that published test sets are typically scarcely relevant for real-world applications, so that this situation may bring in the risk to see the research community around MIR to be "*isolated from real-world interests*" in the near future. This claim sounds as a serious alarm bell for researchers and practitioners of the field. Let us also consider that in concrete scenarios, as the one proposed in (95), the accuracy figures obtained by state-of-the-art tools may not be fully satisfactory for an industrial exploitation (74; 94). Lew et al. give also an interesting hint on some future research directions, including human centered methods, multimedia collaboration, neuroscience methods exploitations, folksonomies.

Integration between semantic Web technologies and multimedia retrieval techniques is considered a future challenge by many researchers (132). In this field, the task of concept detection is concerned with identifying of instances of semantically evocative language terms through the numerical analysis of multimedia items. The work of Bertini et al. (24) proposes a solution in the domain of television sport programmes. Their approach uses a static hierarchy of classes (named pictorially enriched ontology) to describe the prototypical situations findable in football matches and associate them with low-level visual descriptors. In (25), the authors present a complete system for creating multimedia ontologies, automatic annotation, and video sequences retrieval based on ontology reasoning.

8.1.3 A Reference Scenario

Let us consider a family composed by Bob, Alice, and their son Edinson. Bob is an investment broker. Due to his job, he is interested in both economy/finance news and transportation services. Alice is a housewife who cares with the health of her child. Edinson is a sport lover and his favorite sport is football.

Each family member is encountering severe problems for fulfilling personal information

needs. From the Internet point of view, the availability on a daily basis of a wide variety of information sources generates a disproportionately high amount of content (e.g., newspaper articles and news agency releases) that makes it impossible for each of them to read everything that is produced. Furthermore, it is obvious that the Internet is not (yet) the only source of news information, being traditional media based on television channels still far from being left out in the near future. Due to the heterogeneity of individual interests, classical newscast programmes and TV advertising can be extremely inefficient, leaving viewers annoyed and upset.

To accomplish the users' needs, media industry wish would be to have an application able to aggregate data produced from different sources, to give a short description of them in form of keywords, and to associate them with advertising messages according to the content of the generated aggregations. In this way the map is complete and each user can get informed with completeness and efficacy. For example, Bob might stay tuned on last stock market news through his tablet, while being advised on journeys and transports. Alice might watch health care news stories on her television, while being recommended on the healing properties of herbs and natural remedies. Finally, Edinson might browse his favorite team articles, while getting hints on new sport furniture.

8.2 Multimodal Aggregation

Multimodal aggregation of heterogeneous data, also known as *information mash-up*, is a hot topic in the World Wide Web community. A multimodal aggregator is a system that merges content from different data sources (e.g., Web portals, IPTV, etc.) to produce new, hybrid data that was not originally provided. Here, the challenge lies in the ability of combining and presenting heterogeneous data coming from multiple information sources, i.e., *multimedia*, and consisting of multiple types of content, i.e., *cross-modal*. As a result of this technological breakthrough, the content of modern Web is characterized by an impressive growth of multimedia data, together with a strong trend towards integration of different media and modalities of interaction. The mainstream paradigm consists in *bringing into the Web* what was thought (and produced) for different media, like TV content (acquired and published on websites and then made available for indexing, tagging, and browsing). This gives rise to the so-called *Web Sink Effect*. This effect has rapidly started, recently, to unleash an ineluctable evolution from the original concept of the Web as a resource where to *publish* things produced in various forms *outside* the Web, to a world where things *are born and live* on the Web. In this chapter, we adopt Web newspaper articles and TV newscasts as information sources to produce multimodal aggregations of informative content integrating items coming from both contributions. In the following of this section we briefly overview the main ideas behind this task (the interested reader may refer to (97), for further details).

The corresponding system can be thought as a processing machine having two inputs, i.e., digitized broadcast news streams (DTV) and online newspapers feeds (RSSF), and one output, i.e., the multimodal aggregations that are automatically determined from the semantic aggregation of the input streams by applying a co-clustering algorithm whose kernel is an

asymmetric relevance function between information items (98).

Television news items are automatically extracted from the daily programming of several national TV channels. The digital television stream is acquired and partitioned into single programmes. On such programmes, newscast detection and segmentation into elementary news stories is performed. The audio track of each story is finally transcribed by a speech-to-text engine and indexed for storage and retrieval. Further details can be found in (96).

The RSSF stream consists of RSS feeds from several major online newspapers and press agencies. Each published article is downloaded, analyzed, and indexed for search purposes. The first step of the procedure consists in cleaning the downloaded article webpages from boilerplate content, i.e., HTML markups, links, scripts, and styles. Linguistic analysis, i.e., sentence boundary detection, sentence tokenization, word lemmatization and POS tagging, is then performed on the extrapolated contents. The output of this analysis is then used to transform the RSS content into a query to access the audio transcriptions of the DTV news stories, thus allowing to combine text and multimedia in an easy way.

The output of the clustering process is a set of multimodal aggregations of broadcast news stories and newspaper articles related to the same topic. TV news stories and Web newspaper articles are fully cross-referenced and indexed. For each multimodal aggregation, users can use automatically extracted tag clouds, to perform local or Web searches. Local searches can be performed either on the specific aggregation the tags belong to or to the global set of discovered multimodal aggregations. Tag clouds are automatically extracted from each thread topic as follows: (i) each word classified as proper noun by the linguistic analysis is a tag; (ii) a tag belongs to a multimodal aggregation if it is present in at least one aggregated news article; and (iii) the size of a tag is proportional to the cumulative duration of television news items which are semantically relevant to the aggregated news article to which the tag belongs.

Each news aggregation, also called *subject*, is described by a set of attributes, the main being:

- *info*, the general information included title and description;
- *categories*, the set of most relevant categories to which the news aggregation belong. They are automatically assigned by AI:Categorizer[1], trained with radio programme transcriptions, according to a set of journalistic categories (e.g., *Politics*, *Currents Affairs*, *Sports*);
- *tagclouds*, a set of automatically-generated keywords;
- *items*, the set of Web articles that compose the aggregation;
- *videonews*, the collection of relevant newscasts stories that compose the news aggregation.

Hence, a news aggregation is composed by online articles (*items*) and parts of newscasts (*videonews*). In this chapter, we concentrate only in the former. Each item is described as set of attributes, such as:

- *pubdate*, the timestamp of first publication;

[1]http://search.cpan.org/ kwilliams/AI-Categorizer-0.09/lib/AI/Categorizer.pm

- *lastupdate*, the timestamp when the item was updated;
- *link*, the URL of the news webpage;
- *feed*, the RSS feed link that includes the item;
- *title*, the title;
- *description*, the content;
- *category*, the category to which the news belong (according to the previously mentioned classification procedure);
- *keywords*, the keywords automatically extracted as described above.

8.3 Content-based Keyword Extraction to Multimodal News Aggregations

The first aim of this work is to automatically extract keywords to news and news aggregations. In particular, we are aimed at selecting keywords relevant to the news and news aggregations.

Among other solutions, we decided to use suitable extraction-based TS techniques. To this end, we first consider six straightforward but effective extraction-based text summarization techniques proposed and compared in the work of Kolcz (in all cases, a word occurring at least three times in the body of a document is a keyword, while a word occurring at least once in the title of a document is a title-word):

- *Title* (T), the title of a document;
- *First Paragraph* (FP), the first paragraph of a document;
- *First Two Paragraphs* (F2P), the first two paragraphs of a document;
- *First and Last Paragraphs* (FLP), the first and the last paragraphs of a document;
- *Paragraph with most keywords* (MK), the paragraph that has the highest number of keywords;
- *Paragraph with Most Title-words* (MT), the paragraph that has the highest number of title-words.

Let us note that we decided to not consider the Best Sentence technique, i.e. the technique that takes into account sentences in the document that contain at least 3 title-words and at least 4 keywords. This method was defined to extract summaries from textual documents such as articles, scientific papers and books. In fact, news are often inadequate to find meaningful sentences composed by at least 3 title-words and 4 keywords in the same sentence.

Furthermore, we consider the enriched techniques proposed in Chapter 4:

- *Title and First Paragraph* (TFP), the title of a document and its first paragraph:

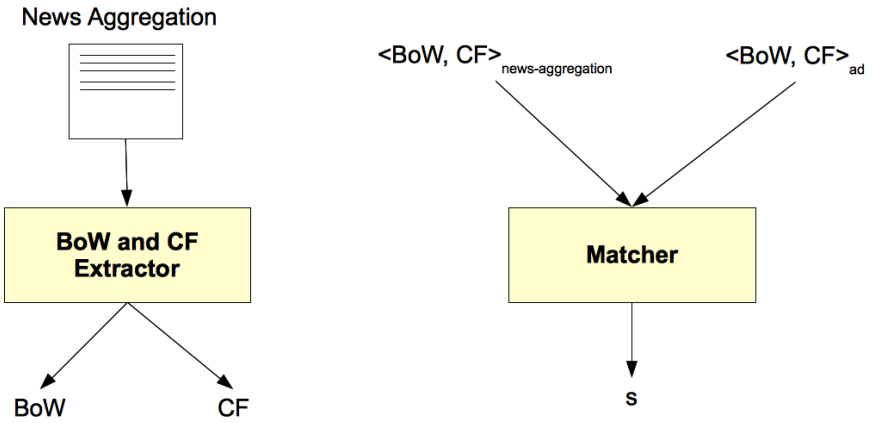

Figure 8.1: The main components of the proposed system. For each news aggregation, both syntactic and semantic information are extracted. Syntactic information is expressed as a bag-of-words (BoW) vector. Semantic information is expressed as a weighted classification feature (CF) vector. The most prominent ads with respect to the news aggregation content are those whose similarity score s are above a given threshold.

- *Title and First Two Paragraphs* (TF2P), the title of a document and its first two paragraphs;
- *Title, First and Last Paragraphs* (TFLP), the title of a document and its first and last paragraphs;
- *Most Title-words and Keywords* (MTK), the paragraph with the highest number of title-words and that with the highest number of keywords.

These techniques have been successfully applied in the CA field, as we showed in Chapter 4.

8.4 Automatic Advertisment Associations to Multimodal News Aggregations

The second aim of the work reported in this Chapter is to automatically suggest relevant advertisements to news and news aggregations. To this end, we developed a suitable CA system.

The proposed system aims to suggest ads that match with the content of a given news aggregation. To this end, we adopt a solution compliant with state-of-the-art CA approaches.

Figure 8.1 illustrates the main components of the prosed system: the *BoW and CF Extractor*, aimed at extracting the Bag-of-Words (BoW) and the Classification Features (CF) of a given news aggregation, and the *Matcher* aimed at selecting the ads according to the similarity with the given news aggregation.

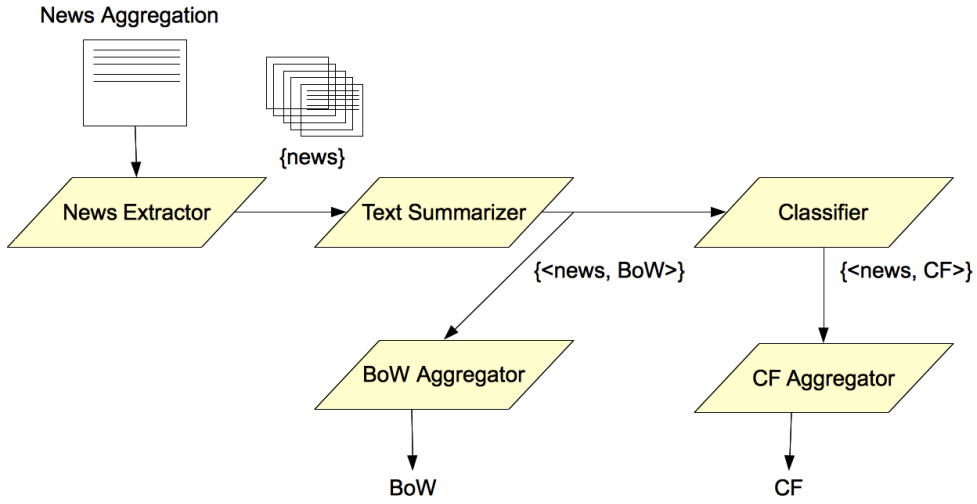

Figure 8.2: The main modules involved in the process of BoW and CF extraction.

Figure 8.2 shows the process of extracting syntactical and semantic features from a news aggregation performed by the BoW vector generator and the CF vector extractor, respectively. Given a news aggregation, the *News Extractor* is aimed at extracting all the news that compose it. In order to transform the news content into an easy-to-process document, any given news is also parsed to remove stop-words, tokenize it and stem each term. Then, for each news, the *Text Summarizer* calculates a vector representation as BoW, each word being weighted by TF-IDF. According to the comparative experiments illustrated in the previous Section, we implemented the text summarization technique that showed the best results, i.e., TF2P. This extraction-based technique takes into account information belonging to the Title and the First Two Paragraphs of the news.

The output of the *Text Summarizer*, i.e., a list of $\langle news, BoW \rangle$ pairs, is given in input to the *BoW Aggregator* that is devoted to calculate the BoW of the whole news aggregation. The aggregated BoW is obtained considering the occurrences in the whole set of news, weighted by TF-IDF. Since, typically, CA systems work with a sole webpage this module is absent in classical CA systems, its goal being to allow us to work with aggregated data. To infer the topics of each news, the *Classifier* analyzes them according to a given set of classes based on a taxonomy of journalistic categories. First, for each class we represent it with its centroid, calculated starting from the training set. We then classify each document by adopting the Rocchio classifier (114) with only positive examples and no relevance feedback. Each centroid component is defined as a sum of TF-IDF values of each term, normalized by the number of documents in the class. The classification is based on the cosine of the angle between the news and the centroid of each class. The score is normalized by the news and class lengths to produce a comparable score. The output of this module is a list of $\langle news, CF \rangle$ pairs, where, in accordance with the work by Broder, CF are the Classification Features extracted by the classifier. The output of the *Classifier*, i.e., a list of $\langle news, CF \rangle$ pairs, is given in input to the *CF Aggregator* that is devoted

to calculate the CF of the whole news aggregation. The aggregated CF is obtained considering the scores giving by the classifier. It is worth noting that, similarly to the *BoW Aggregator*, this module, absent in classical CA systems, allows us to work with aggregated data.

Each ad, which in our work is represented by the webpage of a product or service company, is processed in a similar way and it is represented by suitable BoW and CF. To choose the ads relevant to a news aggregation, the *Matcher* assigns a score s to each ad according to its similarity with a given news:

$$s(n,a) = \alpha \cdot sim_{BoW}(n,a) + (1-\alpha) \cdot sim_{CF}(n,a) \tag{8.1}$$

in which α is a global parameter that permits to control the impact of BoW with respect to CF, whereas $sim_{BoW}(n,a)$ and $sim_{CF}(n,a)$ are cosine similarity scores between the news (n) and the ad (a) using BoW and CF, respectively. For $\alpha = 0$ only semantic analysis is considered, whereas for $\alpha = 1$ only the syntactic analysis is considered.

8.5 Experiments and Results

To assess the effectiveness of the proposed solutions, we perform several experiments. As for the task of extracting keywords, we performed two sets of comparative experiments: (i) experiments on the sole news comparing the performance with those corresponding to the adoption of the keywords provided in the *keyword* attribute and (ii) experiments on news aggregations comparing the performance with those corresponding to the adoption of the keywords provided in the *tagclouds* attribute. Performances have been calculated in terms of precision, recall, and F1 by exploiting a suitable classifier. As for the task of associating suitable ads, we, first, set up a dataset of ads, composed by webpages of product-service companies, and, then, about 15 users[1] were asked to evaluate the relevance of the suggested ads. Performances have been calculated in terms of *precision at k*, i.e., the precision of the system in suggesting k ads, with k varying from 1 to 5.

8.5.1 The Adopted Dataset

As for the comparative study on TS, experiments have been performed on about 45,000 Italian news and 4,800 news aggregations from January 16, 2011 to May 26, 2011. The adopted dataset is composed by XML files, each one describing a subject according to the attributes described in Section 8.2. News and news aggregations were previously classified into 15 categories, i.e., the same categories adopted for describing news and news aggregations.

As for assessing the performances of the proposed CA system, experiments were performed on a set of about 600 news aggregations belonging to the following 8 categories[2]: *Economy and*

[1] Assessors have been selected among students and young researchers of the Department of Electrical and Electronic Engineering as well as workers at RAI Centre of Research and Technological Innovation.

[2] We do not take into considerations categories with a very few numbers of documents and also those that, according to the work of Brother, are not suitable to suggest ads.

Finance (EF), *Environment Nature and Territory* (ENT), *Health and Health Services* (HHS), *Music and Shows* (MS), *Publishing Printing and Mass Media* (PPMM), *Religious Culture* (RC), *Sports* (S), and *Transportation* (T). For each category we selected 11 ads as webpages of product-service companies.

8.5.2 Experimenting Text Summarization

First, we performed experiments on news by adopting a system that takes as input an XML file that contains all the information regarding a news aggregation. For each TS technique, first the system extracts the news, parses each of them, and adopts stop-word removing and stemming. Then, it applies the selected TS technique to extract the corresponding keywords in a vector representation (BoW). To calculate the effectiveness of that technique, the extracted BoW is given as input to a centroid-based classifier, which represents each category with a centroid calculated starting from a suitable training set[1]. A BoW vector is then classified by measuring the distance between it and each centroid, by adopting the cosine similarity measure.

Performances are calculated in terms of precision, recall, and F1. As for the baseline technique (B), we considered the BoW corresponding to the set of keywords of the *keywords* attribute. Table 8.1 summarizes the results.

Table 8.1: Comparisons among TS techniques on news.

	B	**T**	**FP**	**F2P**	**FLP**	**MK**	**MT**	**TFP**	**TF2P**	**TFLP**	**MTK**
P	0.485	0.545	0.625	0.705	0.681	0.669	0.650	0.679	**0.717**	0.706	0.692
R	0.478	0.541	0.594	0.693	0.667	0.665	0.640	0.663	**0.704**	0.697	0.686
F1	0.481	0.543	0.609	0.699	0.674	0.667	0.645	0.671	**0.710**	0.701	0.689
#t	5	5	13	23	22	20	15	16	25	24	18

Subsequently, we performed experiments on news aggregations in a way similar to the one adopted for the sole news. For each TS technique, first the system processes each news belonging to the news aggregation in order to parse it, to disregard stop-words, and to stem each remaining term. Then, it applies to each news the selected TS technique in order to extract the corresponding keywords in a BoW representation. Each extracted BoW is then given in input to the same centroid-based classifier used for the news. The category to which the news aggregation belongs to is then calculated averaging the scores given by the classifier for each involved item.

Table 8.2 shows the results obtained by comparing each TS technique, the baseline (B) being the BoW corresponding to the set of keywords of the *tagclouds* attribute.

Results clearly show that, for both news and news aggregations, TS improves performances with respect to the adoption of the baseline keywords. In particular, best performances in terms

[1]In order to evaluate the effectiveness of the classifier, we performed a preliminary experiment in which news are classified without using TS. The classifier shown a precision of 0.862 and a recall of 0.858.

Table 8.2: Comparisons among TS techniques on news aggregations.

	B	T	FP	F2P	FLP	MK	MT	TFP	TF2P	TFLP	MTK
P	0.624	0.681	0.693	0.764	0.734	0.717	0.727	0.731	**0.770**	0.766	0.737
R	0.587	0.678	0.683	0.766	0.728	0.709	0.718	0.728	**0.769**	0.759	0.729
F1	0.605	0.679	0.688	0.765	0.731	0.713	0.723	0.729	**0.769**	0.762	0.733
#t	62	70	204	319	337	231	200	215	318	338	280

of precision, recall, and –hence– F1, are obtained by adopting the TF2P technique. The last row of Table 8.1 and Table 8.2 shows the number of terms extracted by each TS technique. It is easy to note that, except for the T technique, TS techniques extract a number of terms greater than that extracted by the baseline approach. Let us also note that precision, recall, and F1 calculated for news aggregations are always better than those calculated for news. This is due to the fact that news aggregations are more informative than the sole news and the number of extracted keywords is greater.

8.5.3 Experimenting the Contextual Advertising System

About 15 users were asked to evaluate the relevance of the suggested advertisements and performance were calculated in terms of *precision at* k, i.e., the precision of the system in suggesting k ads, with k varying from 1 to 5.

Preliminary experiments have been performed to calculate the best value of α in Equation 8.1 to maximize the number of correct proposed advertisements. Figure 8.3 shows the results obtained comparing, for each suggested $\langle newsaggregation, ad \rangle$ pair, the category to the news aggregation belongs with the one to the ad belongs, varying α. Results show that the best results are obtained with a value of α in the range $0.25 - 0.40$, meaning that the impact given by the semantic contribution is greater.

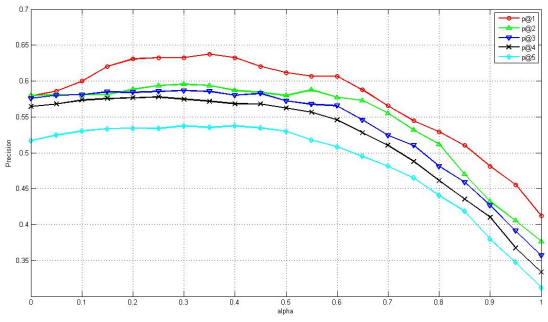

Figure 8.3: *Precision at* k, varying α.

Then, we set α to 0.35 and we asked to the selected assessors to give, for 80 randomly selected news aggregations, a degree of relevance, i.e., relevant (1), somewhat relevant (2), or irrelevant

(3). According to the state of the art, the assessor scores were averaged to produce a composite score and converted in a binary score by assuming as irrelevant $\langle newsaggregation,\ ad \rangle$ pairs with a composite score higher or equal to 2.34. We also calculated the variance, σ_k, and the average value, μ_k, of the assessor agreements, where k is the number of ads suggested by the proposed CA system for each news aggregation.

Table 8.3: Results of the proposed CA system according to the assessor evaluation.

	G	**EF**	**ENT**	**HHS**	**MS**	**PPMM**	**RC**	**S**	**T**
p@1	0.632	0.300	0.600	0.800	0.400	0.667	**0.900**	0.600	0.800
σ_1	0.411	0.197	0.508	0.573	0.359	0.473	0.381	0.386	0.414
μ_1	2.136	2.523	2.319	2.038	2.408	2.128	1.608	2.185	1.882
p@2	0.639	0.450	0.550	0.800	0.450	0.500	**0.900**	0.700	0.750
σ_2	0.429	0.309	0.459	0.566	0.381	0.441	0.410	0.395	0.472
μ_2	2.123	2.420	2.340	1.965	2.342	2.269	1.638	1.984	2.037
p@3	0.594	0.367	0.567	0.733	0.433	0.444	**0.867**	0.633	0.700
σ_3	0.414	0.330	0.431	0.528	0.374	0.405	0.385	0.367	0.492
μ_3	2.154	2.470	2.331	2.026	2.308	2.339	1.654	2.046	2.080
p@4	0.573	0.375	0.550	0.725	0.425	0.389	**0.850**	0.625	0.625
σ_4	0.400	0.338	0.428	0.520	0.340	0.367	0.368	0.361	0.477
μ_4	2.179	2.452	2.322	2.044	2.335	2.400	1.665	2.077	2.158
p@5	0.542	0.380	0.540	**0.720**	0.420	0.378	0.680	0.580	0.620
σ_5	0.391	0.331	0.432	0.528	0.330	0.342	0.342	0.362	0.459
μ_5	2.224	2.450	2.300	2.051	2.362	2.458	1.882	2.141	2.170

Table 8.3 reports the results in terms of $p@k$, σ_k, and μ_k. Those results show that the $p@k$ is on average around 0.6. This is due to several issues. First, news aggregation descriptions are often too short and not enough informative for the assessor. Furthermore, let us note that some noise might be introduced by the fully automatic process of aggregation building. Moreover, some categories are very specific (e.g., *Religious Culture*) whereas others are very generic (e.g., *Economy and Finance*). To put into evidence this issue, the following columns in Table 8.3 (from **EF** to **T**) show the performances of the system in terms of $p@k$, σ_k, and μ_k, for each category.

Results clearly show that the more the category is specific, the better the system performs. To better highlight this point, Figure 8.4 shows the variance (σ_1) and the average value (μ_1) of assessor agreements considering only one advertisement for each category, i.e. setting $k = 1$. The green point is the average value μ_1 of the scores given by the assessors; whereas the red line shows the variance σ_1 around μ_1. Let us stress the fact that, for the sake of readability, we chose to display σ_1 around μ_1 just to give a visual hint. The Figure shows that the shorter the red line is, the greater the agreement is. For instance, for *Economy and Finance*, assessors agree

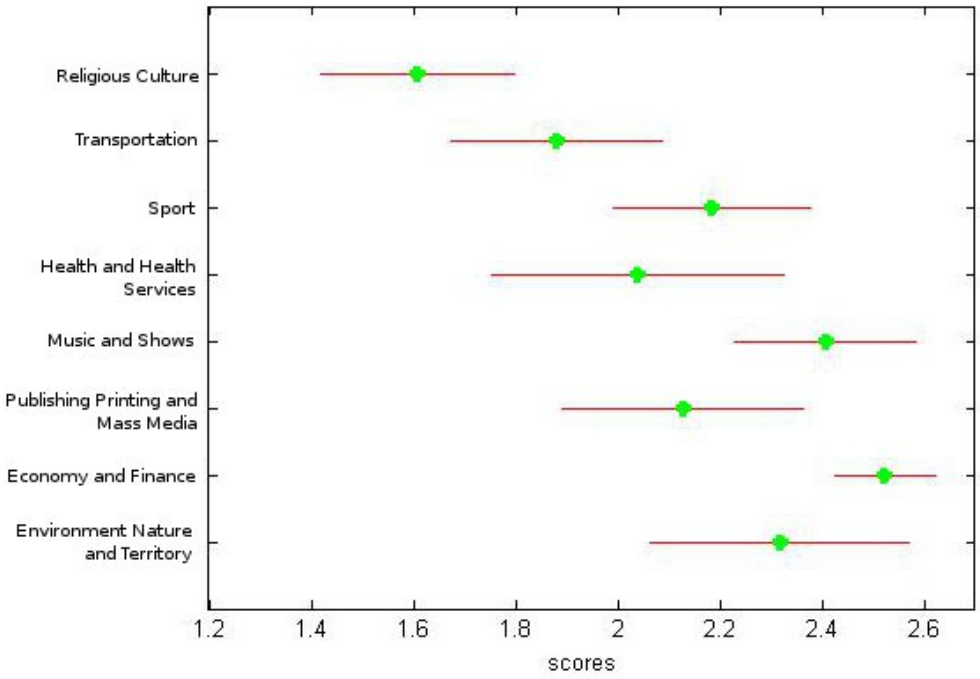

Figure 8.4: The variance (σ_1) and the average value (μ_1) of assessor agreements for each category and for $k = 1$.

that the suggested ads are mostly *irrelevant*. On the other hand, for *Health and Health Services*, even if the average value is 2.04 (i.e., *somewhat relevant*), assessors are in disagreement.

8.6 The Proposed Approach in the Reference Scenario

To better illustrate how the proposed solutions work in practice, let us go back to the reference scenario sketched in Section 8.1.3. Suppose that Edinson is searching for information regarding the football match "Real Madrid - Barcelona" and he finds the news aggregation whose description is shown in Figure 8.5[1].

"Merengues" recovered after a sending off, but distance from the leadership of "Blaugrana" remains of 8 points. The goals of Messi and Cristiano Ronaldo, both by penalty kick, opened the row of "clasicos", in Liga, Copa del Rey, and Champions.
The final score of 1-1 delivers the Spanish championship to "Blaugrana", which holds a distance of 8 points from the rivals, but they failed to maintain the row of 5 successive wins of "clasicos", although the asset of one player and one goal. "Merengues" tied with a proud response, useless for winning the Liga, but that morally can be considered as a victory, since it happened after a row of defeats and bad performances.

Figure 8.5: An example of description of a news aggregation.

According to our proposal solution, the system analyzes all the news that compose the aggregation (in this case 42) and then each of them is summarized according to the TF2P technique and classified by the centroid-based classifier. All the BoW and the CF are collected by the *BoW Aggregator* and the *CF Aggregator*, respectively. Thanks to this approach, to evaluate if reading all the aggregated news, Edinson could first consider the suggested keywords shown in Figure 8.6.

champions
Barcelona
supporters
Ronaldo
extratime
Blaugrana
team
triumph
Liga
Madrid
Real
goal
Mourinho
catalan

Figure 8.6: A selection of the keywords extracted from the news aggregation in Figure 8.5.

Then, if Edinson decides to read the news, the system calculates, for each ad in the ads repository, the matching with the given news aggregation according to the equation 8.1. In so doing, the system proposes to him the 5 ads depicted in Figure 8.7. As shown in the Figure,

[1]Currently, we are considering only Italian news, for the sake of clarity we translated the news in English.

Figure 8.7: The suggested ads.

all the suggested ads are related to the main category *Sport*: a kit to synchronize your runner shoes with your mp3 player; two online sport newspapers; a website that sells sport and show tickets; and a video-game that allows users to make sports.

8. CONTEXTUAL ADVERTISING ON MULTIMODAL AGGREGATION

Chapter 9

Conclusions and Future Directions

9.1 Conclusions

In this thesis novel solutions in the field of CA have been studied, developed, and evaluated.

The first part of the work has been the focusing on the syntactic task of the advertising process, with the analyses and the development of improved text summarization techniques. In particular, we first proposed some straightforward extraction-based techniques by taking into account the information of the title of a webpage. Experimental results confirm the hypothesis that adding such information to well-known techniques allows to improve performances. Then, we proposed to adopt snippet as summaries of a webpage. The comparative study showed that the proposed snippet text summarization technique has similar performances, in terms of precision, recall, and F_1 with respect to the proposed extractive technique. The impact of each summarization technique has been studied and evaluated in a CA system. Experimental results confirmed the intuition that the adoption of enriched extractive techniques improves the performances of the classic methods, in particular by taking into account the information of webpage title. Furthermore, the impact of snippets has also been studied in the field of CA. The adoption of snippets as text summarization technique in CA showed that performances, calculated in terms of precision at k, are quite good, especially in suggesting 1 or 2 ads and that the system that uses both snippets and title is the one with the best performances.

The subsequently research activity has been focused on the study of the semantic phase in the process of advertising with particular reference to the extraction of classification-based features. In particular, we proposed to adopt a semantic enrichment by taking into account the concepts extracted by a lexical resource named ConceptNet. To this end, we devised a system called ConCA (Concepts in Contextual Advertising). The system has been evaluated in terms of precision by performing comparative experiments with a state-of-the-art system. Results shown that the adoption of concepts positively affects the choice of ads.

9. CONCLUSIONS AND FUTURE DIRECTIONS

In the following, we presented a unifying view of CA and RS. We proposed a hybrid CA system that exploits collaborative filtering in a content-based setting. Collaborative filtering here is intended in that, to suggest an ad to a given webpage, we exploit the collaboration of suitable pages related to it, i.e., pages similar to it, at least, in the topics. The peer pages in our vision could be represented by the related links. To this end, we conducted a preliminary experimental study to investigate the impact of related links in a generic CA system. In particular, given a webpage p, as related links we chose a set of inlinks (i.e., the pages that link to p) and a set of outlinks (i.e., the pages that are linked by p). To study the impact of collaborative filtering in CA, we performed comparative experiments with a content-based system that does not resort to any collaborative approach. Experiments have been performed considering the impact of related links alone and in conjunction with the original page. Our results show that the adoption of related links improves of about 9.2% the performance of the baseline system that rely only on the content of the given page. So that we can sentence that related links are effective for CA. To better highlight how the system works and its effectiveness, we also provided a suitable case study, i.e., how to suggest ads to the webpage of a webpage. Furthermore, we proposed a novel RS devised *a la mode of* CA. The proposed system is able to recommend new items to users according to their profile, which in turn is represented by a set of categories extracted from a reference taxonomy. Experiments, performed using WNDH, highlight that the proposed approach is quite effective, also considering that it has been adopted to suggest photos with a scarce description.

Another branch of the research activity is related to the conjunct study of CA and geo-location. We investigated the background and propose a preliminary study on the users' behaviors in the field of mobile services. In particular, we analyzed the correlation between the user location and the location of the service linked by the clicked URL. The users are often more interested in services with a local intent, and it's possible characterize a subset of users with that intents.

Finally, we investigated the application of CA in the field of multimodal aggregation. Multimodal aggregation is one of the most emerging topics in the World Wide Web research field. In this challenging scenario, information engineering and integration plays a vital role in optimizing costs and quality of the provided services. We proposed a preliminary solution that applies classical CA solutions to suggest advertisements to aggregations of news stories from television and the Internet. As for each aggregation, the included news stories are fully cross-referenced, advertisements can be automatically placed in both the webpages linked by the RSS articles and the TV news stories associated to them. The first research task has been a preliminary study aimed at verifying the effectiveness of adopting text summarization techniques to suggest keywords to news and news aggregations in a multimodal aggregation system. To perform our study, we compared several different extraction-based techniques with the keywords provided by the adopted multimodal aggregation system. Results, calculated in terms of precision, recall, and F1, shown that the best set of keywords is obtained considering the title, the first and second paragraph of each news. Then we proposed a CA system, compliant with state-of-the-art approaches, that has been experimented on a test set of news aggregations. To calculate the

overall performances, we made several experiments aimed at measuring the precision at k, i.e., the precision of the system in suggesting k ads, with k from 1 to 5. Performances have been calculated in two different settings: (i) measuring precision according to previously classified news and ads and (ii) asking to a set of users to give a degree of relevance. Results showed that the system reach good performances in terms of precision. As expected, the system performs very well with some specific categories. On the contrary, poor performances are obtained for more generic categories, mainly due to the heterogeneity of their contents.

9.2 Future directions

As for future directions, we are currently setting up and studying several methods to improve all the proposed solutions.

Further experiments are also under way. In particular, we are planning to evaluate the text summarization techniques by performing further comparative experiments with the methods described in the literature and with different and larger datasets.

We are studying novel semantic techniques. The main idea is to improve the known techniques by exploiting semantic information extracted from different lexical resources (e.g., WordNet or Yago) Further experiments are also under way. In particular, we are setting up the proposed systems to calculate its performances with a larger dataset extracted by DMOZ in which documents are categorized according to a given taxonomy of classes.

In the same direction, further possible research is the adoption of a hierarchical text categorization, instead of the modified Rocchio classifier. We deem that taking into account the taxonomic relationship among classes would improve the overall performance of the classifier.

As for the future work of the collaborative filtering system, we are setting up further experiments to investigate the impact of related links that take into account the title or the URL of each page, in conjunction with its snippet. Further techniques to select similar pages are under study, as for example link prediction methods and the adoption of complex networks or clustering techniques.

A further direction, although it requires considerable efforts, is to implement user profiling to personalize the selected ads according to user's tastes and preferences.

As for the research on Multimodal Aggregation we are currently implementing a prototype of the system described in this work, able to be integrated in the HyperMedia News, a system for the automated aggregation and presentation of information streams from digital television and the Internet. This is a jointed work with the RAI Centre for Research and Technological Innovation.

References

[1] S. Abbar, M. Bouzeghoub, and S. Lopez. **Context-Aware Recommender Systems: A Service-Oriented Approach**. In *3rd International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases*, 2009. 58

[2] A. Addis, E. Angioni, G. Armano, R. Demontis, F. Tuveri, and E. Vargiu. **A Novel Semantic Approach to Document Collections**. *IADIS International Journal on Computer Science and Information Systems*, **4(2)**, 2009. 75

[3] Andrea Addis, Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu. **A Novel Recommender System Inspired by Contextual Advertising Approach**. In *Proceedings of IADIS'10: International Conference Intelligent Systems and Agents*, pages 67–74, 2010. 72

[4] Andrea Addis, Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu. **A Recommender System based on a Generic Contextual Advertising Approach**. In *Proceedings of ISCC'10: IEEE Symposium on Computers and Communications*, pages 859–861, 2010. 72

[5] Andrea Addis, Giuliano Armano, and Eloisa Vargiu. **Profiling Users to Perform Contextual Advertising**. In *Proceedings of the 10th Workshop dagli Oggetti agli Agenti (WOA 2009)*, 2009. 73

[6] Gediminas Adomavicius and Alexander Tuzhilin. **Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions**. *IEEE Transactions on Knowledge and Data Engineering*, **17**(6):734–749, 2005. 55, 57

[7] Gediminas Adomavicius and Alexander Tuzhilin. **Context-aware recommender systems**. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 335–336, New York, NY, USA, 2008. ACM. 58

[8] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. **Web-a-where: geotagging web content**, 2004. 79

[9] Aris Anagnostopoulos, Andrei Z. Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. **Just-in-time contextual advertising**. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 331–340, New York, NY, USA, 2007. ACM. 16

[10] S.S. Anand and B. Mobasher. *Contextual Recommendation*, pages 142–160. Springer, Germany, 2007. 58

[11] Chinatsu Aone, Scott William Bennett, and Jim Gorlinsky. **Multi-Media Fusion through Application of Machine Learning and NLP**. In *In AAAI Spring Symposium Working Notes on Machine Learning in Information Access*, 1996. 9

[12] Douglas E. Appelt. **Introduction to information extraction**. *AI Commun.*, **12**:161–172, August 1999. 9

[13] G. Armano, M. de Gemmis, G. Semeraro, and E. Vargiu. *Intelligent Information Access*, **SCI 301**. Springer-Verlag, Studies in Computational Intelligence series, Heidelberg, Germany, 2010. 55

[14] Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu. **Experimenting Text Summarization Techniques for Contextual Advertising**. In *IIR'11: Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop*, 2011. 33

[15] Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu. **Semantic Enrichment of Contextual Advertising by Using Concepts**. In *International Conference on Knowledge Discovery and Information Retrieval*, 2011. 49

[16] Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu. **Studying the Impact of Text Summarization on Contextual Advertising**. In *8th International Workshop on Text-based Information Retrieval*, 2011. 38

[17] Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu. **Using Snippets in Text Summarization: a Comparative Study and an Application**. In *IIR'12: 3rd Italian Information Retrieval (IIR) Workshop*, 2012. 35

[18] Giuliano Armano and Eloisa Vargiu. **A Unifying View of Contextual Advertising and Recommender Systems**. In *Proceedings of International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010)*, pages 463–466, 2010. 55, 59

[19] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. **Spatial variation in search engine queries**. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 357–366, New York, NY, USA, 2008. ACM. 80

[20] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999. 1, 30

[21] Ziv Bar-Yossef and Maxim Gurevich. **Random sampling from a search engine's index**. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 367–376, New York, NY, USA, 2006. ACM. 5

[22] R. K. Belew. *Finding out about: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge University Press, 2000. 30

[23] Michael W. Berry. *Survey of Text Mining*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003. 8

[24] M. Bertini, A. Del Bimbo, and C. Torniai. **Enhanced ontologies for video annotation and retrieval**. In *Proc. of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 89–96, 2005. 89

REFERENCES

[25] M. Bertini, A. Del Bimbo, and C. Torniai. **Automatic annotation and semantic retrieval of video sequences using multimedia ontologies**. In *Proc. of the 14th annual ACM international conference on Multimedia*, pages 679–682, 2006. 89

[26] Krishna Bharat and Andrei Broder. **A technique for measuring the relative size and overlap of public Web search engines**. *Comput. Netw. ISDN Syst.*, **30**:379–388, April 1998. 5

[27] L. Boratto and S. Carta. *State-of-the-Art in Group Recommendation and New Approaches for Automatic Identification of Groups*, pages 1–20. Soro A., Vargiu E., Armano G. and Paddeu G. (Eds.), Springer, Germany, 2010. 57

[28] Ronald Brandow, Karl Mitze, and Lisa F. Rau. **Automatic condensation of electronic publications by sentence selection**. *Inf. Process. Manage.*, **31**:675–685, September 1995. 31

[29] Sergey Brin and Lawrence Page. **The anatomy of a large-scale hypertextual Web search engine**. *Comput. Netw. ISDN Syst.*, **30**:107–117, April 1998. 60

[30] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. **A semantic approach to contextual advertising**. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 559–566, New York, NY, USA, 2007. ACM. 13

[31] Alexander Budanitsky and Graeme Hirst. **Evaluating WordNet-based Measures of Lexical Semantic Relatedness**. *Comput. Linguist.*, **32**:13–47, March 2006. 60

[32] R. Burke. **Hybrid Recommender Systems: Survey and Experiments**. *User Modeling and User-Adapted Interaction*, **12**(4):331–370, 2002. 57

[33] Mary Elaine Califf and Raymond J. Mooney. **Relational learning of pattern-match rules for information extraction**. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, AAAI '99/IAAI '99, pages 328–334, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence. 9

[34] J. Carrasco, D. Fain, K. Lang, and L. Zhukov. **Clustering of bipartite advertiser-keyword graph**. In *Proc. International Conference on Data Mining (ICDM'03)*, Melbourne, Florida, November 2003. 13

[35] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. **Focused crawling: a new approach to topic-specific Web resource discovery**. *Computer Networks (Amsterdam, Netherlands: 1999)*, **31**(11–16):1623–1640, 1999. 60

[36] Patrali Chatterjee, Donna L. Hoffman, and Thomas P. Novak. **Modeling the Clickstream: Implications for Web-Based Advertising Efforts**. *Marketing Science*, **22**:520–541, October 2003. 14

[37] Peter Cheeseman, James Kelly, Matthew Self, John Stutz, Will Taylor, and Don Freeman. *Readings in knowledge acquisition and learning*, chapter AutoClass: a Bayesian classification system, pages 431–441. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. 9

[38] Massimiliano Ciaramita, Vanessa Murdock, and Vassilis Plachouras. **Online learning from click data for sponsored search**. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 227–236, New York, NY, USA, 2008. ACM. 16

[39] Massimiliano Ciaramita, Vanessa Murdock, and Vassilis Plachouras. **Semantic Associations for Contextual Advertising**. *Journal of Electronic Commerce Research*, **9(1)**:1–15, 2008. Special Issue on Online Advertising and Sponsored Search. 13

[40] Paul R. Cohen and Rick Kjeldsen. **Information retrieval by constrained spreading activation in semantic networks**. *Information Processing and Management*, **23**(4):255–268, 1987. 60

[41] Dipanjan Das and André F.T. Martins. **A Survey on Automatic Text Summarization**. Technical Report Literature Survey for the Language and Statistics II course at CMU, 2007. 30

[42] K. de Smedt, A. Liseth, M. Hassel, and H. Dalianis. *How short is good? An evaluation of automatic summarization*, pages 267–287. Museum Tusculanums Forlag, Kbenhavn, 2005. 30

[43] Chakrabarti Deepayan, Agarwal Deepak, and Josifovski Vanja. **Contextual advertising by combining relevance with click feedback**. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2008. ACM. 72

[44] Koen Deschacht and Marie-Francine Moens. **Finding the Best Picture: Cross-Media Retrieval of Content**. In *Proc. of ECIR 2008*, pages 539–546, 2008. 88

[45] N. Dimitrova. **Multimedia Content Analysis: The Next Wave.** In *CIVR'03 Proceedings of the 2nd international conference on Image and video retrieval*, pages 9–18, 2003. 89

[46] Junyan Ding, Luis Gravano, Narayanan Shivakumar, and Gigabeat Inc. **Computing Geographical Scopes of Web Resources**. pages 545–556, 2000. 79

[47] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. **Inductive learning algorithms and representations for text categorization**. In *Proceedings of the seventh international conference on Information and knowledge management*, CIKM '98, pages 148–155, New York, NY, USA, 1998. ACM. 30

[48] H. P. Edmundson. **New Methods in Automatic Extracting**. *Journal of ACM*, **16**:264–285, April 1969. 30

[49] Daniel C. Fain and Jan O. Pedersen. **Sponsored search: A brief history**. *Bul. Am. Soc. Info. Sci. Tech.*, **32**(2):12–13, 2006. 12

[50] Douglas H. Fisher. **Knowledge Acquisition Via Incremental Conceptual Clustering**. *Mach. Learn.*, **2**:139–172, September 1987. 9

[51] H. P. Frei and D. Stieger. **The use of semantic links in hypertext information retrieval**. *Inf. Process. Manage.*, **31**:1–13, January 1995. 60

[52] Dayne Freitag. **Machine Learning for Information Extraction in Informal Domains**. *Mach. Learn.*, **39**:169–202, May 2000. 9

[53] Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. **CACTUS: clustering categorical data using summaries**. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 73–83, New York, NY, USA, 1999. ACM. 30

[54] Claire Green and Peter Edwards. **Using Machine Learning to Enhance Software Tools for Internet Information Management**. In *Proceedings of the AAAI Workshop on Internetbased Information Systems*, pages 48–55, 1996. 9

[55] Eui-Hong Han and George Karypis. **Centroid-Based Document Classification: Analysis and Experimental Results**. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '00, pages 424–431, London, UK, 2000. Springer-Verlag. 22

[56] U. Hanani, B. Shapira, and P. Shoval. **Information Filtering: Overview of Issues, Research and Systems**. *User Modeling and User-Adapted Interaction*, **11**:203–259, 2001. v, 6

[57] A. Hanjalic. *Content-Based Analysis of Digital Video*. Kluwer Academic Publishers, December 2004. 89

[58] Catherine Havasi, Robert Speer, and Jason Alonso. **ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge**. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September 2007. 48

[59] S. He and K. Tanaka. **Modeling Omni-Directional Video**. In *Advances in Multimedia Modeling, 13th International Multimedia Modeling Conference, MMM 2007*, pages 176–187, 2007. 89

[60] Steffen Heinz and Justin Zobel. **Efficient single-pass index construction for text databases**. *J. Am. Soc. Inf. Sci. Technol.*, **54**:713–729, June 2003. 2

[61] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. **On near-uniform URL sampling**. *Comput. Netw.*, **33**:295–308, June 2000. 5

[62] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. **Recommending and evaluating choices in a virtual community of use**. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co. 57

[63] Scott B. Huffman. **Learning information extraction patterns from examples**. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 246–260, London, UK, 1996. Springer-Verlag. 9

[64] A. Jameson and B. Smyth. *Recommendation to Groups*, pages 596–627. Brusilovsky P., Kobsa A., and Nejdl W. (Eds.), Springer, Germany, 2007. 57

[65] T. Joachims. **Optimizing Search Engines Using Clickthrough Data**. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002. 13

[66] Sue J. Ker and Jen-Nan Chen. **A text categorization based on summarization technique**. In *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11*, pages 79–83, Morristown, NJ, USA, 2000. Association for Computational Linguistics. 30

[67] Svetlana Kiritchenko. *Hierarchical text categorization and its application to bioinformatics*. PhD thesis, Univ. of Ottawa, Canada, Ottawa, Ont., Canada, Canada, 2006. 8

[68] Jon M. Kleinberg. **Authoritative sources in a hyperlinked environment**. *Journal of ACM*, **46**:604–632, September 1999. 60

[69] M. Kokar. **Formalizing classes of information fusion systems**. *Information Fusion*, **5**(3):189–202, September 2004. 88

[70] Aleksander Kołcz and Joshua Alspector. **Asymmetric Missing-data Problems: Overcoming the Lack of Negative Data in Preference Ranking**. *Inf. Retr.*, **5**:5–40, January 2002. 31

[71] Aleksander Kolcz, Vidya Prabakarmurthi, and Jugal Kalita. **Summarization as feature selection for text categorization**. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 365–370, New York, NY, USA, 2001. ACM. 31

[72] Daphne Koller and Mehran Sahami. **Hierarchically classifying documents using very few words**. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 170–178, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US. 8

[73] Marijn Koolen and Jaap Kamps. **Are semantically related links more effective for retrieval?** In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 92–103, Berlin, Heidelberg, 2011. Springer-Verlag. 60

[74] W. Kraaj, A. Smeaton, and P. Over. **TRECVID 2004: An Overview**. In *Proc. of TRECVID Workshop 2004*, 2004. 89

[75] Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, and Berthier Ribeiro-Neto. **Learning to advertise**. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 549–556, New York, NY, USA, 2006. ACM. 15

[76] C. Laudy and J.-G. Ganascia. **Information fusion in a TV program recommendation system**. In *11th International Conference on Information Fusion, 2008*, pages 1–8, 2008. 88

REFERENCES

[77] Steve Lawrence and C. Lee Giles. **Searching the world wide web**. *Science*, **280**(5360):98–100, 1998. 5

[78] R. Lempel and S. Moran. **SALSA: the stochastic approach for link-structure analysis**. *ACM Transactions on Information Systems*, **19**:131–160, April 2001. 60

[79] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. **Content-Based Multimedia Information Retrieval: State of the Art and Challenges**. *ACM Transactions on Multimedia Computing, Communications and Applications*, **2**(1):1–19, 2006. 89

[80] X. Li, J. Yan, Z. Deng, L. Ji, W. Fan, B. Zhang, and Z. Chen. **A novel clustering-based RSS aggregator**. In *Proc. of WWW07*, pages 1309–1310, 2007. 88

[81] Ray Liere and Prasad Tadepalli. **The Use of Active Learning in Text Categorization**. In *Association for the Advancement of Artificial Intelligence*, pages 591–596, 1997. 9

[82] G. Linden, B. Smith, and J. York. **Amazon.com Recommendations**. *IEEE Internet Computing*, **07**(1):76–80, 2003. 57

[83] H. Liu and P. Singh. **ConceptNet: A Practical Commonsense Reasoning Tool-Kit**. *BT Technology Journal*, **22**:211–226, October 2004. 48

[84] P. Lops, M. de Gemmis, and G. Semeraro. *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. Ricci, F., Rokach, L., Shapira, B. ans Kantor, P.B. (Eds.), Springer, US, 2010. 57

[85] H.P. Luhn. **The automatic creation of literature abstracts**. *IBM Journal of Research and Development*, **2**:159–165, 1958. 30

[86] B. Magnini and G. Cavagli. **Integrating Subject Field Codes into WordNet**. In *Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.)* Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, pages 1413–1418, 2000. 74

[87] R. P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House, Inc., Norwood, MA, USA, 2007. 88

[88] I. Mani. *Automatic summarization*. John Benjamins, Amsterdam, 2001. 31

[89] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. v, 1, 3, 4

[90] Massimo Marchiori. **The quest for correct information on the Web: hyper search engines**. *Comput. Netw. ISDN Syst.*, **29**:1225–1235, September 1997. 60

[91] M. E. Maron and J. L. Kuhns. **On Relevance, Probabilistic Indexing and Information Retrieval**. *Journal of ACM*, **7**:216–244, 1960. 7

[92] Joel Martin. **Clustering Full Text Documents**. In *Proceedings of the IJCAI Workshop on Data Engineering for Inductive Learning at IJCAI-95*, 1995. 9

[93] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. **A probabilistic approach to spatiotemporal theme pattern mining on weblogs**. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 533–542, New York, NY, USA, 2006. ACM. 80

[94] A. Messina, W. Bailer, P. Schallauer, and V. Tablan et al. **Content Analysis Tools**. Deliverable 15.4, PrestoSpace Consortium, February 2007. 89

[95] A. Messina, L. Boch, G. Dimino, W. Allasia, and R. Basili et al. **Creating rich Metadata in the TV Broadcast Archives Environment: the PrestoSpace project**. In *Proc. of IEEE AXMEDIS06 Conference*, 2006. 89

[96] A. Messina, R. Borgotallo, G. Dimino, D. Airola Gnota, and L. Boch. **ANTS: A Complete System for Automatic News Programme Annotation based on Multimodal Analysis**. In *Intl. Workshop on Image Analysis for Multimedia Interactive Services*, 2008. 91

[97] Alberto Messina and Maurizio Montagnuolo. *Information Retrieval and Mining in Distributed Environments*, chapter Multimodal Aggregation and Recommendation Technologies Applied to Informative Content Distribution and Retrieval. A. Soro and E. Vargiu and G. Armano and G. Paddeu, 2010. 90

[98] Alberto Messina and Maurizio Montagnuolo. **Heterogeneous Data Co-Clustering by Pseudo-Semantic Affinity Functions**. In *Proc. of the 2nd Italian Information Retrieval Workshop (IIR)*, 2011. 91

[99] George A. Miller. **WordNet: A Lexical Database for English**. *Commun. ACM*, **38**(11):39–41, 1995. 74

[100] Dunja Mladenić and Marko Grobelnik. **Feature Selection for Classification Based on Text Hierarchy**. In *Text and the Web, Conference on Automated Learning and Discovery CONALD-98*, 1998. 30

[101] Alistair Moffat and Timothy A. H. Bell. **In situ generation of compressed inverted files**. *J. Am. Soc. Inf. Sci.*, **46**:537–550, August 1995. 2

[102] Vanessa Murdock, Massimiliano Ciaramita, and Vassilis Plachouras. **A noisy-channel approach to contextual advertising**. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, ADKDD '07, pages 21–27, New York, NY, USA, 2007. ACM. 5

[103] L.-D. Nguyen, K.-Y. Woon, and A.-H. Tan. **A self-organizing neural model for multimedia information fusion**. In *11th International Conference on Information Fusion, 2008*, pages 1–7, 2008. 88

[104] Justin Picard and Jacques Savoy. **Enhancing retrieval with hyperlinks: a general model based on propositional argumentation systems**. *J. Am. Soc. Inf. Sci. Technol.*, **54**:347–355, February 2003. 60

[105] M. Porter. **An algorithm for suffix stripping**. *Program*, **14**(3):130–137, 1980. 63

[106] D. R. Radev. **A common theory of information fusion from multiple text sources step one: cross-document structure**. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue*, pages 74–83, Morristown, NJ, USA, 2000. Association for Computational Linguistics. 88

[107] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. **Introduction to the special issue on summarization**. *Computational Linguistic*, **28**:399–408, December 2002. 30

[108] L. Ramaswamy, R. Polavarapu, K. Gunasekera, D. Garg, K. Visweswariah, and S. Kalyanaraman. **Caesar: A Context-Aware, Social Recommender System for Low-End Mobile Devices**. In *IEEE International Conference on Mobile Data Management*, pages 338–347, 2009. 58

[109] L. F. Rau, P. S. Jacobs, and U. Zernik. **Information extraction and text summarization using linguistic knowledge acquisition**. *Inf. Process. Manage.*, **25**:419–428, June 1989. 30

[110] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. **GroupLens: an open architecture for collaborative filtering of netnews**. In *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, New York, NY, USA, 1994. ACM. 57

[111] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher, and Edleno Silva de Moura. **Impedance coupling in content-targeted advertising**. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 496–503, New York, NY, USA, 2005. ACM. 15

[112] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor. *Recommender Systems Handbook*. Springer, US, 2010. 30, 56

[113] Elaine Rich. *User modeling via stereotypes*, pages 329–342. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998. 57

[114] J. Rocchio. *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval, pages 313–323. PrenticeHall, 1971. 33, 44, 63, 94

[115] Paat Rusmevichientong, David M. Pennock, Steve Lawrence, and C. Lee Giles. **Methods for Sampling Pages Uniformly from the World Wide Web**. In *In AAAI Fall Symposium on Using Uncertainty Within Computation*, pages 121–128, 2001. 5

[116] G. Salton and C. Buckley. **On the use of spreading activation methods in automatic information**. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '88, pages 147–160, New York, NY, USA, 1988. ACM. 30

[117] Gerard Salton. *A theory of indexing*. Philadelphia : Society for Industrial and Applied Mathematics, 1975. 1

[118] Gerard Salton and Michael McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984. 20

[119] Badrul Sarwar, George Karypis, Joseph Konstan, and John Reidl. **Item-based collaborative filtering recommendation algorithms**. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM. 57, 76

[120] F. Sebastiani. **Machine Learning in Automated Text Categorization**. *ACM Computing Surveys (CSUR)*, **34**(1):1–55, 2002. 7, 8

[121] Azadeh Shakery and Cheng Xiang Zhai. **A probabilistic relevance propagation model for hypertext retrieval**. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 550–558, New York, NY, USA, 2006. ACM. 60

[122] Upendra Shardanand and Pattie Maes. **Social information filtering: algorithms for automating "word of mouth"**. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 210–217, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co. 57

[123] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, and Wei-Ying Ma. **Web-page classification through summarization**. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 242–249, New York, NY, USA, 2004. ACM. 30

[124] Mark Sinka and David Corne. **A Large Benchmark Dataset for Web Document Clustering**. In *Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications*, pages 881–890. Press, 2002. 24

[125] C. G. Snoek and M. Worring. **Multimodal video indexing: A review of the state-of-the-art**. In *Multimedia Tools and Applications*, pages 5–35, 2005. 89

[126] Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. **CRYSTAL inducing a conceptual dictionary**. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pages 1314–1319, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 9

[127] K. Stefanidis and E. Pitoura. **Fast Contextual Preference Scoring of Database Tuples**. In *EDBT*, pages 344–355, 2008. 58

[128] Frank Curtis Stevens. *Knowledge-based assistance for accessing large, poorly structured information spaces*. PhD thesis, Boulder, CO, USA, 1993. 57

[129] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. **PHOAKS: a system for sharing recommendations**. *Communication of ACM*, **40**(3):59–62, 1997. 57

REFERENCES

[130] Srinivas Vadrevu, Ya Zhang, Belle L. Tseng, Gordon Sun, and Xin Li. **Identifying regional sensitive queries in web search.** In Jinpeng Huai, Robin Chen, Hsiao-Wuen Hon, Yunhao Liu, Wei-Ying Ma, Andrew Tomkins, and Xiaodong Zhang, editors, *WWW*, pages 1185–1186. ACM, 2008. 80

[131] Chong Wang, Jinggang Wang, Xing Xie, and Wei-Ying Ma. **Mining geographic knowledge using location aware topic model**. In Ross Purves and Chris Jones, editors, *Proceedings of the 4th ACM Workshop On Geographic Information Retrieval, GIR 2007, Lisbon, Portugal, November 9, 2007*, pages 65–70. ACM, 2007. 80

[132] T. Wang, N. Yu, Z. Li, and M. Li. **nReader: reading news quickly, deeply and vividly**. In *In Proc. of CHI '06 extended abstracts on Human factors in computing systems*, pages 1385–1390, 2006. 89

[133] Y.Z. Wei, N.R. Jennings, L. Moreau, and W. Hall. **User Evaluation of a Market-based Recommender System**. *Autonomous Agents and Multi-Agent Systems*, **17**:251–268, 2000. 7

[134] Peter Willett. **Recent trends in hierarchic document clustering: a critical review**. *Inf. Process. Manage.*, **24**:577–597, August 1988. 9

[135] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Com- pressing and Indexing Documents and Images*. Morgan Kaufmann, 1999. 2, 30

[136] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. **Optimal multimodal fusion for multimedia data analysis**. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579, New York, NY, USA, 2004. ACM. 88

[137] C. Xu, J. Wang, H. Lu, and Y. Zhang. **A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video**. *IEEE Trans. on Multimedia*, **10**(3):421–436, 2008. 88

[138] Tak W. Yan and Hector Garcia-Molina. **The SIFT information dissemination system**. *ACM Transactions on Database Systems*, **24**(4):529–565, 1999. 57

[139] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. **Finding advertising keywords on web pages**. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 213–222, New York, NY, USA, 2006. ACM. 15

[140] Ziming Zhuang, Cliff Brunk, and C. Lee Giles. **Modeling and visualizing geo-sensitive queries based on user clicks**. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, pages 73–76, New York, NY, USA, 2008. ACM. 80

[141] Justin Zobel and Alistair Moffat. **Inverted files for text search engines**. *ACM Comput. Surv.*, **38**, July 2006. 2

[142] Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. **On assigning place names to geography related web pages**. In *JCDL '05: Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 354–362, New York, NY, USA, 2005. ACM Press. 79