Alma Mater Studiorum · University of Bologna

SCHOOL OF SCIENCE

Master Degree in Computer Science

Temporal PageRank

Supervisor:
Prof. Danilo Montesi

Candidate:
Luca Pompei

Session III
Academic Year 2014 - 2015

To my family, my love, Arianna,
my friends and to all the people who have accompanied and supported me during these wounderful years.

Sommario

La realtà moderna, il cui sviluppo incessante ne rappresenta uno degli aspetti principali, è caratterizzata da un'evoluzione costante e decisamente importante. Per seguire questa crescita e questo progresso, tuttavia, le persone necessitano di rimanere sempre aggiornate sugli avvenimenti, accaduti e che accadono, alimentando così il loro fabbisogno informativo.

In un mondo in cui, però, le notizie esistenti sono rappresentate da un numero indiscutibilmente elevato, ricercare quali siano quelle corrette, ideali o semplicemente vicine a quelle maggiormente attese e desiderate può costituire un problema assolutamente considerevole e da non sottovalutare, considerando che gli ostacoli che rendono arduo tale compito si amplieranno sempre più nel tempo, a causa dell'arricchimento dei dati a nostra disposizione.

Un grande aiuto, in risposta a queste incertezze e preoccupazioni, è dato dall'*Information Retrieval*, una nuova e interessante branca interdisciplinare dell'informatica che si occupa della gestione e del recupero delle giuste informazioni. Sulla base di un dataset di riferimento, più o meno ampio e contenente al suo interno un insieme di documenti eterogenei dal punto di vista dei formati che li descrivono, un sistema di Information Retrieval è sviluppato al fine di ricercare e restituire contenuti ritenuti rilevanti sulla base di un'esigenza esprimibile attraverso una query interrogativa che gli utenti possono fornire.

Al fine di soddisfare al meglio tali ambizioni e aspirazioni, bisogna render noto che la

maggior parte dei sistemi elaborati in questo settore e conosciuti allo stato dell'arte fanno affidamento esclusivamente sulla similarità testuale per identificare le informazioni rilevanti, definendole tali nel caso in cui comprendano la presenza di una o più keywords manifestate dalla query utilizzata come rappresentante del bisogno informativo.

L'idea qui portata avanti e studiata, e già sostenuta dalla comunità scientifica, è che tutto ciò non sempre basta e risulta sufficiente, in particolar modo quando bisogna gestire database dalle grandi dimensioni, come lo è il mondo del web. Le migliori soluzioni esistenti, infatti, potrebbero, in questi casi, generare risposte e risultati di bassa qualità e non permettere all'utente interessato una valida navigazione attraverso essi. L'intuizione, per superare queste limitazioni, è stata quella di definire un nuovo concetto di rilevanza, grazie al quale ordinare diversamente le pagine web ritenute più soddisfacenti.

Concentrandosi su queste motivazioni, di conseguenza, è stata data luce a **Temporal PageRank**, una nuova proposta per il *Web Information Retrieval* che fa affidamento sulla combinazione di diversi fattori per tentare di aumentare la qualità di una ricerca sul web.

Temporal PageRank incorpora, in primo luogo, i vantaggi di un classico algoritmo di ranking, grazie al quale vengono privilegiate le informazioni riportate dalle pagine web ritenute importanti dal contesto stesso in cui risiedono, e, in secondo luogo, le potenzialità di tecniche appartenenti al mondo del *Temporal Information Retrieval*, in maniera tale da sfruttare gli aspetti temporali dei dati racchiusi in esse, descriventi i loro contesti cronologici di appartenenza.

In questa tesi, quindi, la nuova proposta è descritta e trattata dettagliatamente, confrontando i risultati riportati da essa con quelli raggiunti dalle migliori soluzioni conosciute allo stato dell'arte, analizzando quindi pregi e difetti che lo caratterizzano.

Abstract

The modern reality, whose incessant development represents one of the main aspects, is characterized by a constant and decisively important evolution. To follow this growth and this progress, however, people need to stay up to date on the events, that took place and that happen, thus feeding their information need.
In a world in which, anyway, the existing news are represented by an unquestionably high number, search for the correct and ideal ones, or simply for those that are close to the more expectated and desired, may constitute an absolutely considerable and not to be underestimated problem, considering that the obstacles that make arduous this task will be expanded more and more over time, due to the enrichment of the data at our disposal.
A great help, in response to these concerns and uncertainties, is given by *Information Retrieval*, an exciting new interdisciplinary branch of computer science that deals with the management and the retrieval of the right information. On the basis of a reference dataset, more or less wide and containing inside it a set of documents heterogeneous from the point of view of the formats that describe them, an Information Retrieval system is developed in order to return and search for contents considered relevant on the basis of a need expressed by an interrogative query that users can provide.
In order to satisfy these ambitions and aspirations, we must make known that most of the systems developed in this area and known to the state of the art, rely solely on textual similarity to identify relevant information, defining them as such in the case

that these include the presence of one or more keywords expressed by the query used as representing of the informative need.

The idea expressed and studied here, and already supported by the scientific community, is that all that is not always enough and sufficient, especially when it is necessary to manage large databases, as is the world wide web. The best existing solutions may, in these cases, generate low quality responses and results and they may not allow, to the interested user, a valid navigation through them. The intuition, to overcome these limitations, has been to define a new concept of relevance, thanks to which differently rank the web pages deemed more satisfactory.

By focusing on these reasons, consequently, the light was given to **Temporal PageRank**, a new proposal for the *Web Information Retrieval* that relies on a combination of several factors for groped to increase the quality of research on the web.

Temporal PageRank incorporates, in a first place, the advantages of a classic ranking algorithm, thanks to which prefer the information reported by those web pages considered important by the context itself in which they reside, and, in a second place, the potential of those techniques that take part to the world of the *Temporal Information Retrieval*, so as to exploit the temporal aspects of data contained in them, describing their chronological contexts of belonging.

In this thesis, then, the new proposal is described and discussed in detail, comparing the results reported by it with those achieved by the best solutions known to the state of the art, analyzing, so, the strengths and the weaknesses that characterize it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Science is built of facts the way a house is built of bricks; but an accumulation of facts is no more a science than a pile of bricks is a house."

— Henri Poincaré

Nowadays, we live in a world in which knowledge, learning and the need to get more and more information are vital. Our daily lives, in fact, moves us in this direction, but, despite this incessant and growing desire, the search for the right news is not in any case simple and trivial. This primary requirement is commonly found by us in a multitude of fields, each one different from the other, such as in the professional scope, or in the personal one or else in many others, only to mention a few.
The complexity of a search varies from situation to situation and the main factors not to be underestimated are, first of all, the size that characterizes the dataset in which there is the need to undertake this operation and, of course, the type and the format of data contained inside it. The difficulty in spotting specific information, among the many available, furthermore, whether they are represented by numbers, text, images, audio or other else, undoubtedly depends on their nature, on their design but also on the possibility that these collections can include inconsistent and conflicting contents

between them, or even data that are no longer usable because they have lost their validity or utility over time.

Although the many technological tools at our disposal today, indeed, under many points of view, these elements represent some of the biggest obstacles during the search for the most relevant and suitable documents for our purposes.

The most important example, that we can refer to, is without doubts the web, a dynamic pool of information that has reached incredible numbers, extremely vast and rapidly growing, as evidenced by many studies about it. Thinking of the universe of the web, entering in detail, it is equivalent to identify a huge warehouse of data possibly different between them, continuously changing in time and stored inside of structures not necessarily created and maintained in the same way. The amount of data that it can have, among the many difficulties that one can encounter, can make the search particularly costly, both in terms of time taken to successfully conclude them and in terms of precision in finding information, among the many available, that best fit the requests received by interested people.

To satisfy our needs in front of a world with very extensive dimensions, therefore, we must research the correct information with care and with the right methods, identifying those we like more and those that are closest to the expected and waited ideal solution.

In this first chapter, so, the background on which my thesis work is introduced, indicating the reasons that, finally, led me to the formulation and the identification of a new research method to retrieve information on the web, considering all the possible problems in which we may encounter and the possible solutions thanks to which properly overcome them.

1.1 Background

The search in a big dataset, as already mentioned, is a complex task that must be done with the right commitment, the right attention and the right tools.
Almost many years have now passed since the idea of J.E.Holmstrom [21] which envisaged to use, for the first time, a machine in order to search for desired information. From that moment on, this area of research has attracted more and more interest and it has developed a lot, until the formation of the branch of computer science that is now defined as **Information Retrieval (IR)** and that involves various application fields such as computer science, information architecture, design, linguistics, semiotics, information science, cognitive psychology, and even more.
The meaning of the latter term is very broad, and it refers to the activity or else to the entire set of techniques focused on the search and the retrieve of relevant information derived from a need of them, expressed in a textual form or in other nature.
The Information Retrieval, in practice, is an interdisciplinary field that deals with the representation, the storage and the organization of textual information. To reach its goal, it tries to return a set of appropriate *objects* with respect to an initial *query*. The query in question is nothing more than a real interrogation to the dataset, consisting of a string of keywords and it represents the required information that we need to recover. The objects, instead, correspond to the collection of those results that have been returned with the aim to satisfy our informative need.
There are two main ways to assess the quality of the latter and understand, at the same time, how much the retrieved and recovered information are valid and inherent with respect to the initial analyzed request and they are, respectively, the precision and recall property:

$$precision = \frac{relevantDocuments \cap retrievedDocuments}{retrievedDocuments}$$

$$recall = \frac{relevantDocuments \cap retrievedDocuments}{relevantDocuments}$$

The first property corresponds to the proportion of relevant documents among those recovered, while the second one is equivalent to the proportion between the number of relevant retrieved documents and the number of all relevant available documents.
One of the most important challenges for the Information Retrieval, among many, is the analysis of collections of unstructured data and documents, organized, so, in an heterogeneous way between them.
In this sense, it is enough think about the world of the web.
Various approaches and several methods have been developed to address these issues, in the context of the IR identified as **Web Information Retrieval (Web-IR)**. Nowadays, it is enough consider the role played by search engines, both those commercial and not.

1.2 Motivation

In vast datasets, as is the web but how many others could be, we cannot now limit to the simple search of news on the basis of the **textual similarity** that they have and they reach in accordance with the expressed informative desire. Many of the results thus found, in fact, may not be relevant or suitable with respect to the initial request, such as in cases where, to mention a few, the latter represents a very generic information, or, in those situations in which so many answers may be returned, making so impossible a valid user navigation through these.

Wanting to make an example, suppose we are interested in the main remedies against influenza and, therefore, we assume that the user intending to operate this research formulates, on the web, a very generic query like "main remedies against influenza", using a particular search engine that relies solely on textual similarity to satisfy the questions that are asked to it. It is easy to deduce that, as a result of this demand, it is difficult, for the user, completely and properly analyze all the possible outcomes that could be returned, since they could be a really huge number.

What has been obtained cannot certainly be considered satisfactory, beyond the values of precision and recall that the method has gained, observing that explore such a large amount of data is challenging and may lead to the achievement of no objective, which, as previously mentioned, is to search for the main remedies against influenza.

We must try to give a new sense of "relevance" to better evaluate and rank the documents and the web pages.

To improve the quality of the research is possible to intervene in several ways, by acting exploiting different techniques. First of all, we can resort to classical **ranking algorithms**, thanks to which sort the list of solutions offered by the search engine in question, by placing in the top positions those that are considered the most reliable and suitable elements in accordance with the user requirements or those coming from authoritative sources, such as may be medical journals, government sites or many others, and, at the same time, discarding or relegating in the last positions of the ranking those deemed not useful or else not relevant to the problem to be solved.

This is not enough. The use of classical ranking algorithms is necessary, but not sufficient to provide an ideal solution in every circumstance. We must explore new ways to identify and assess, in a better way, the relevance of the researched information, in order to better satisfy the informative needs of users.

To clarify what said above, we assume now another possible scenario in which there is the need to find out the name of the President of the United States of America. In

order to know this information, we could hypothesize a generic query on the web like "President of the United States of America", but, in doing so, a search engine might return the list of all those who over time have held that position, creating so a set of results including many unnecessary information, not helpful to the resolution of the question posed above.
The idea, to overcome these limitations, is to put into consideration every **aspect of time**, concerning both the asked question and the retrieved news pertaining to the web or, generalizing, pertaining to any other large dataset.
These problems should not be overlooked, since they will become ever larger, important and accident over time, with the increase of the datasets, because the growth that characterizes the web, as noted, is and will be unstoppable.

1.3 Problem statement

One of the most interesting challenges for the Information Retrieval is come across in the retrieve of relevant documents belonging to large databases, as is the world of the web.
The latter, in fact, is undoubtedly an information repository for knowledge reference of huge size, rapidly growing, containing data written in different formats, different languages and, above all, time-varying. These factors, therefore, make the search of news an extremely complex operation, for which, consequently, it is no longer enough to recover and display them. The real difficulty, in fact, is to show them in the most effective and relevant way with respect to the informative need that the user has expressed, rewarding those data that seem to be most suitable and best reflect the initial desire that we must satisfy, and, viceversa, penalizing other ones that do not contain the right features and are less useful.

We must look, as anticipated, for new solutions and new approaches in order to help users in their researches, taking into account not only the textual similarity existing between demand and response, but also considerating, in addiction to the page content and to any other information available through it, the graphical structure that regulates the web world.
The idea behind this thesis, to overcome these obstacles and to introduce a new sense of relevance to judge the objects to be searched, is to exploit, in a first place, a ranking algorithm to reward and penalize, according to certain criteria, the elements contained in the set of the retrieved results and, in a second time, to consider the all existing temporal aspects contained both in queries and in web pages.
The goal to achieve with this work, concluding, is to contribute, furthermore, to increase the quality of the final answer during a search on the web.

1.4 Organization

After treating the introduction in this chapter, it is below described the manner in which this thesis is organized.
In the next chapter, to continue, the state of the art is discussed, analyzing the main currently available proposals, involving, therefore, the problems and the solutions that have alternated in time regarding the Web Information Retrieval. In doing so, the light has been given to the most significant aspects that were considered in the solution presented here, highlighting how they can help to improve those that represent the current state of the art.
In the chapter 3, then, *Temporal PageRank* is introduced, analyzing the principles and the reasons on which it is based, the choices that were taken about it for its design and its construction and the objectives that it arises and that it tried to reach.

In the chapter 4, moving forward, the studies that have been carried on *Temporal PageRank* are considered, focusing especially on how these have been applied and the scenarios in which this proposal is tested and challenged. It is described the reference dataset on which the search operations have been carried out and the impact that each internal factor to *Temporal PageRank* caused.

In chapter 5, thereafter, the obtained goals are shown and discussed, highlighting the cases in which *Temporal PageRank* can not be effective and viceversa. For each behavior that it assumes, therefore, it is provied an explanation in relation to its configuration.

In chapter 6, finally, on the basis of what previously stated, the space was left to the most important conclusions about the study carried out, the results found in relation to the goals that had been set and additional arguments that may push even more the research to improve the actual proposal. In this sense, in fact, several suggestions are offered for possible future works, indicating those that may be the best roads to take to improve the present proposal.

Chapter 2

State of the art

"There is ... a machine called the Univac ... whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape. By this means the text of a document, preceded by its subject code symbol, can be recorded ... the machine ... automatically selects and types out those references which have been coded in any desired way at a rate of 120 words a minute."

— J. E. Holmstrom, 1948

The problems that afflict and make complex search on the web are lots. Over time, however, some of these have been dealt very well, in several ways and with different techniques, trying, when possible, to improve those which, from time to time, were the best known solutions at the state of the art.
In doing so, consequently, thanks to the numerous works offered by the scientific community, we have made considerable progress as regards to the Information Retrieval and the reality of the web, but, despite this, many other issues remained still unresolved, and therefore should be addressed and resolved in order to contribute to the improvement of existing proposals in this area, ensuring that the tools made in support of the Web Information Retrieval will be more efficient and useful for all users and not

only.
This chapter, therefore, deals with the discussion of the main solutions related to the state of the art taken into account for the performance of my studies and the development of this proposal in the context of the Web-IR.

2.1 Okapi BM25

After defining a dataset in which there are the documents or the web page on which operate the researches, as widely known by the scientific community, there are different ways for groped to satisfy the demand placed by the informative desire of users.
The idea to easier fulfill this task and search for the expected news is to identify which of these documents, or web pages, includes inside it at least one of the keywords expressed by the used query.
The methods that combine these concepts are a lot, developed and evaluated in numerous ways.
Okapi BM25 is, in its application domain, the best known solution at the state of the art, both for the world of the web and not, concerning the recovery of information. It is a probabilistic model developed by Stephen E. Robertson et al. [20], thanks to which retrieve the documents, or the web pages, relying solely on the **textual similarity** obtained from them in relation to the query used to interrogate the dataset.
Consequently, this definition makes it a measure dependent on the query submitted to the search engine, which will return, then, different results from time to time, in relation to the considered keywords.
The Okapi BM25 model, therefore, can be defined as a recovery function that sorts the entire set of documents, present in the dataset in which they are searched, according to the observed recurrences, in each of them, of the different terms expressed in the

query, without taking into account the possible existence or not of a relation between these latter.

Let Q be the query used to interrogate the web and let D be the document retrieved as a result of this operation, the score obtained for it by the model in question is mathematically defined as:

$$BM25(Q,D) = \sum_{i=1}^{n} IDF(q_i) * \frac{f(q_i,D) * (k_1+1)}{f(q_i,D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})}$$

in which:

- $\boldsymbol{q_i}$ corresponds to the i_{th} term of the query Q.
- $\boldsymbol{f(q_i, D)}$ is equivalent to the frequency of q_i within the document D that is being evaluated.
- $\boldsymbol{|D|}$ is the length of the document in words.
- ***avgdl*** is the average document length in the text collection from which documents are drawn.
- $\boldsymbol{k_1}$ and $\boldsymbol{b}$ correspond to free parameters, usually chosen as $k_1 \in [1.2, 2.0]$ and $b = 0.75$.

- ***IDF*(q_i)** is the inverse document frequency weight of the query term q_i. It is usually computed as:

$$IDF(q_i) = log * \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.
There are several interpretations for *IDF* and slight variations on its formula. In the original Okapi BM25 derivation, the *IDF* component is derived from the Binary Independence Model.

Several studies and numerous tests were performed, by the scientific community, to evaluate the method in question, as in [19]. Many researchers, in fact, have tried to develop its variants in order to try to improve the results found through it, both as regards to the web and as regards to other generic datasets.
Even now, however, the Okapi BM25 method continues to be considered the central reference for the Information Retrieval, concerning the pure textual similarity existing between queries and documents, or web pages, in order to compare new systems and new approaches in relation to the problem of data recovery.

2.1.1 Implementation

For the following thesis work, the Okapi BM25 model, adopted and used to evaluate each web page or document belonging to any dataset, was applied through

Terrier [1] (version 4.1) [2], the note platform of Information Retrieval, developed at the School of Computing Science, University of Glasgow [3].
Terrier, the search engine, is written in Java and fully implements the Okapi BM25 model, in accordance with the needs expressed above. Its usage is very simple and its customization is equally easy, with the ability to edit each of its main parameters.
The probabilistic model, in its original version, is described through a Java class called *BM25* and it is below provided a discussion about some of its most significant methods and aspects.

Type	**Method signature**
double	getParameter()
double	score(double tf, double docLength)
void	setParameter(double b)

Table 2.1: Main methods of the BM25 implementation in Terrier.

- **double score(double tf, double docLength)**: it uses the Okapi BM25 formula to compute a weight for a term in a document, returning a double value. The first parameter specifies the term frequency in the document, while the second one specifies the document's length.

- **double getParameter()**: it returns a double value corresponding to the b parameter of the Okapi BM25 ranking formula.

[1] Accessible via `http://www.terrier.org/`
[2] Accessible via `http://terrier.org/download/`
[3] Accessible via `http://www.gla.ac.uk/schools/computing/`

- **void setParameter(double b)**: it sets the b parameter used by the Okapi BM25 ranking formula.

In this Okapi BM25 implementation, with regard to the default values, the k_1 parameter is set equal to 1.2, while the b parameter is set equal to 0.75, as suggested by Stephen E. Robertson et al. in their original formulation.

2.1.2 Sample application

Suppose we are interested in using the Okapi BM25 method, through the implementation just presented, in order to satisfy the informative needs of a particular user, by introducing, for this purpose, a test scenario characterized by a decidedly small and trivial size, on which to apply it. The latter, in fact, is constituted by only four documents, whose contents are reported in the table below:

Document	Content
A	*"I play football every day"*
B	*"This is not my green house"*
C	*"My beatiful house"*
D	*"Football is the most popular sport in Italy"*

Table 2.2: Test scenario and web pages beloging to it.

Imagining, now, to exploit two simple queries to interrogate the database defined

above, constituted, each one of them, by only two keywords:

Query	Content
N.1	*"Play football"*
N.2	*"Beatiful house"*

Table 2.3: Queries used to analyze the test scenario.

Without entering again into the merits of details characterizing the model in question, the results obtained from the submission of the two queries on the database of reference are shown below.
As the two tables point out, in both cases, the model reward the documents that include and cover, within their content, the informative desire represented, respectively, by the two adopted queries.

Query	Document	Ranking
N.1	A	(1^o)
	D	(2^o)
	B	(3^o)
	C	(4^o)

Table 2.4: Ranking obtained using the query N.1 through the Okapi BM25 implementation on the test scenario.

Query	Document	Ranking
N.2	C	(1^o)
	B	(2^o)
	A	(3^o)
	D	(4^o)

Table 2.5: Ranking obtained using the query N.2 through the Okapi BM25 implementation on the test scenario.

In the first test case, so, as a result of the query *N.1*, the most relevant document becomes the one identified by the letter A, followed respectively by the letter D, with B and C that appear the lastest in this special classification.
In the second experiment, however, in relation to the query *N.2*, the document most suitable, that is closest to the expected solution by the user, is the one identified by the letter C, followed, quickly after it, by the letter B, until arriving to the letters A and D, representing the solutions that are not suitable and useful for this purpose.
In any case, this banal case can easily be extended into a more sophisticated scenario, applying the previously treated concepts.

2.2 PageRank

One of the fundamental aspects and roles in the Web-IR, which arouse a particular interest in the scientific community and not only, is played by **ranking algorithms**, decisively important, thanks to which is possible to prioritize, in various ways, the obtained and recovered relevant results, in such a way as to increase the quality of the

searches in the World Wide Web.

Among the major and most efficient ranking algorithms for the Web Information Retrieval there is, without doubts, PageRank, created by Sergey Brin and Lawrence Page and shown in 1998 in [7], during the presentation of the system Google [4]. As admitted by the same company inventor in 2008, the Vice President of that time, Udi Manber, asserted that the PageRank method is *"the most important part of the positioning algorithm in Google"* [5], and it continues to be, as even now claimed by other sources and other studies concerning its comprehension.

The intuition of the two authors, confirmed by their analysis and not only, is that the search for relevant information based solely on keywords matching returns, in general, low quality results, especially when the amount of data to manage is very high. In this regard, they introduce a new concept of "ranking" for the web pages, with the aim to obtain qualitatively better results.

The simple, but powerful, idea that is at its base is to exploit the existing citations, or rather the hyperlinks (links), between the several sites that populate the universe of the web. Through them, the ranking algorithm gives life to a map, a *citations graph* through which confer a measure of objective importance to a web page, thanks to which summarize the subjective opinions of importance expressible by people relatively to the same.

PageRank, so, proves to be independent with respect to the query submitted to the search engine and, consequently, each web page will keep the same PageRank value varying the used queries.

According to these concepts, then, the two authors expressed their intention to reward and penalize the pages belonging to the web on the only basis of their incoming and outgoing links, in order to identify the most authoritative sources of information (the

[4] Accessible via `http://www.google.com`

[5] Accessible via Google's official blog at `https://goo.gl/iqmYnS`

most linked or the ones linked from pages considered as such) or less.
Deepening this, the working principle of the PageRank method, then, expect to count the links coming from the different pages, and then to normalize them through the number of links existing on those pages that mention the site that is being analyzed, attributing to them, so, right weights and relevances. Mathematically, therefore, taking a page P that has n pages that pointing to it, named $P_1, P_2, \ldots P_n$ (citations), each with a number of outbound links manifested with $C(P_i)$, and considering a damping factor d, usually contained in the range [0,1], the calculation of its PageRank is so expressible:

$$PR(P) = (1 - d) + d(\frac{PR(P_1)}{C(P_1)} + \ldots + \frac{PR(P_n)}{C(P_n)})$$

The damping factor d mentioned above is intended to simulate a "random surfer" on the web, which, starting from a particular page that is currently visiting, it gets bored and decides to change that page, moving on a new one through the hyperlinks existing on the first. The probability with which the new web page is selected, then, is equal to the value of its *link-based* PageRank score. According to the tests carried out by the two authors of the algorithm, in the article in which it is announced, the optimal value that allows to achieve better results, giving the right power and the most suitable popularity to the web sites, is equal to 0.85.
The damping factor d can be applied both to a single page and, otherwise, to a whole group or set of them, as normally occurs, permitting, in case of its alteration, a new customization, useful in the cases in which there is the need to reward or penalize a particular source of information on the web, contrasting its authority.
PageRank, so, corresponds to a probability distribution over the web pages and, as proof of this, the sum of the all PageRank values earned by the entire set examined

through it is always exactly equivalent to 1.

The purpose of this method, to conclude, is that of groped to imitate human behavior and assess and understand, consequently, the values of the results found by a search, by classifying them in the most appropriate manner to ensure the highest quality during a search on the web world.

The only PageRank, however, has some limitations and the same Google uses other techniques to support it, but these have never been treated in public and no one precisely knows what they are and how they were applied.

Nevertheless, in relation to PageRank, some very good works had been done to try to clarify its properties and its features, evaluating it in detail.

For example, [4] looks inside the method, introducing an analysis that tries to explain the distribution of the page score. The theoretical properties of PageRank, indeed, are only partially understood, but, despite this, a lot of people cite the general theory of Markov chain [6] at its base. The latter work, so, deals with: the issue of its computation, concerning its stability, its complexity and the critical role of involved parameters, providing, in addiction, a more general interpretation of PageRank in the case in which the graph changes over time; the role played by each page without outbound links, named *dangling page*; the interaction amongst communities (group of pages) and their promotion. The latter argument is very interesting, considering that the web visibility can be improved working both on page content and also on pattern of connections. On the other hand, in fact, one can promote a site not only relying on the topological structure of the given community, but in addiction by exploiting external links coming from other communities (this phenomenon is commonly known as "spamming technique").

A different and significant contribute was given in [14]. It corresponds to an extension of the previous work, a survey of all issues associated with PageRank, such as stor-

[6] Accessible via `https://en.wikipedia.org/wiki/Markov_chain`

age issues, existence, uniqueness and convergence properties, possible alterations to the basic model, suggested alternatives to the traditional solution methods, sensitivity and conditioning, and finally the updating problem.
A special attention to the dangling page, however, it has been paid in [13].
Another important work for the study of PageRank is [11]. To yield more accurate search results, it proposes, computing a set of PageRank vectors, biased using a set of representative topics, to capture, more accurately, the notion of importance with respect to a particular topic. In doing so, then, the method allows the query to influence the link-based score. The set of PageRank vectors is computed offline, each biased with a different topic, to create, for each page, a set of importance scores with respect to some particular topics. All this because, in their idea, pages considered important in some subject domains may not be considered important in others.
Finally, thanks to Richardson et al. [18], another important study was conducted about PageRank. According to the latter, it is possible to significantly improve the computational and the efficiency of the method that is being analyzed, through the use of some characteristics that are independent in relation to the structure of the web. On the base of this hypothesis, they designed RankNet, a modified version of PageRank, which uses a machine learning ranking algorithm to assess the frequency with which users visit certain web pages. In doing so, RankNet found encouraging results.

2.2.1 Implementation

The PageRank method can be calculated and applied in different modes, with implementions which can be more or less equivalent. It is possible, for example, think to detect it by a simple iterative algorithm or, otherwise, it is possible to resort to a more sophisticated and efficient alternative in terms of time, developing an its own version of parallel computing, and so on for other new and different ideas.

As there is no a unique officially recognized version of the PageRank method, the chosen implemention for it, however, was sought through the state of the art, so as to have an already tested, functional and known (by the scientific community) version.
Before the final choice of the latter, it has been compared with other versions available through the state of the art, using some test datasets, to verify its effectiveness and its functionality.
The adopted algorithm, so, is written in Python and its (open) source code is available at *GitHub* [7] at the following web address: `https://github.com/kedar-phadtare/PageRank`. It is specifically developed to evaluate the importance of a web page within the collection of data which is used in this thesis to analyze the results reported by the proposal presented here.
Below, the main iterative process that calculates the PageRank scores for the different web pages belonging to the used dataset is discussed and summarized.

Step	**Operation**
1	Initial reading of dataset
2	First assignment of the PageRank scores
3	Iterative calculation of PageRank
4	Convergence and assigning of the final PageRank values

Table 2.6: Main steps of the PageRank computation inside the chosen implementation.

- **Initial reading of dataset**: after reading the reference dataset, the algorithm becomes aware of all the nodes belonging to the entire collection of data to be analyzed and of all the links (and of all their directions) existing inside it, defining so:

[7] Accessible via `https://www.github.com`

- **P** as the set of web pages, of size N.

- **S** as the set of web pages that do not contain outbound links, the so-called *dangling page*, that must be differently managed.

- $\boldsymbol{M(P_i)}$ as the set of web pages with links directed to the generic web page P_i.

- $\boldsymbol{L(P_i)}$ as the number of outgoing links from the generic web page P_i.

- **d** as the damping factor, which can still be customized, set to *0.85*, as suggested by S. Brin and L. Page in the presentation of their system.

- **First assignment of the PageRank scores**: known the nodes corresponding to the different web pages, it assigns an initial PageRank score to each of them equal to *1/N*, with N the total number of nodes. In this way the sum of the all obtained PageRank values is equal to 1.

- **Iterative calculation of PageRank**: iteratively, starting from the values defined above, it calculates the PageRank values for each considered web page, updating them, of course, from time to time.

- **Convergence and assigning of the final PageRank values**: after the convergence of the obtained PageRank values, the algorithm returns the list of nodes sorted on the base of their gained PageRank scores. The convergence is achieved

only in the case in which the same results are reached for 4 consecutive times for each of the studied nodes. When this situation happens, the algorithm stops its execution. However, the parameter that indicates the convergence can be modified.

- **List of results**: as default, the model implementation shows the only first 10 web pages, or else those that appear to be the most important, but, nevertheless, this parameter can easily be changed in the source code, according to the needs that we may have.

2.2.2 Sample application

Considering the original PageRank version and, then, the adopted implementation, a first example of what said, to better clarify this new situation, is visible in the table below and in the figure that accompanies it.
On the basis of the test scenario announced before in Table 2.2, suppose, now, that between the web pages belonging to the dataset are present some links, summarized by the following *citations map* created through the PageRank method:

Web page	**Outbound links**	**Inbound links**
A	B, C	B
B	A	A
C	D	A, D
D	C	C

Table 2.7: Citations map of the test scenario.

As we can imagine, then, the web pages more linked by hyperlinks, or rather those that possess a lot of web sites that direct their links to them, reach higher PageRank scores and, therefore, have a greater importance. At the same time, unlike the previous case, a page can also gain importance in the circumstances in which the web pages that point to it have an high PageRank value.
Please, consider now the figure below.

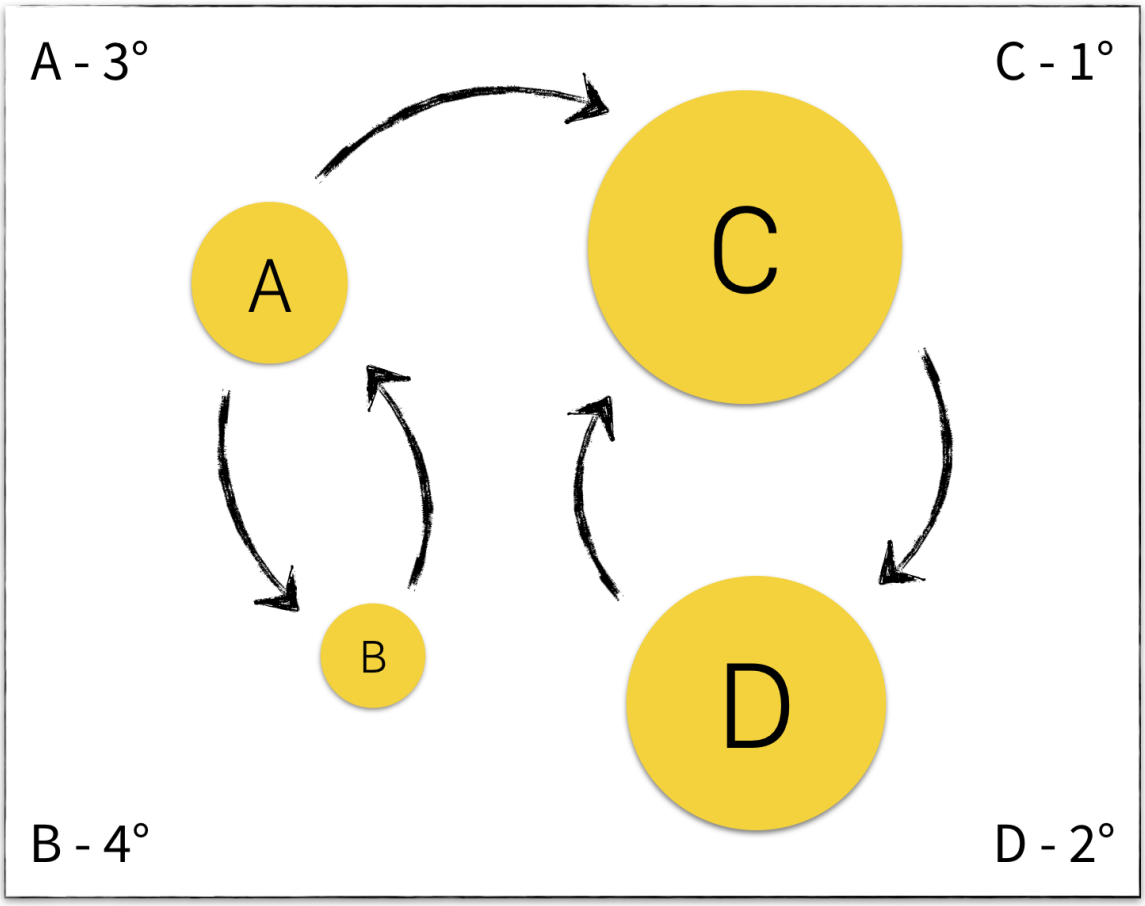


Figure 2.1: Example of PageRank application in a scenario with 4 web pages. In the corners are indicated, for each page, the positions reached in the ranking.

In the image used here, as observable, the circles of greater circumference correspond to web sites of greater importance, while the circles of smaller circumference are related to less relevant web pages. Each arrow, however, constitutes an one-directional link, which start from a source pointing to a new online destination. In the scenario in question, the page that assumes a greater relevance is the one identified by the letter *C*, possessing a single outbound link (directed to *D*) and being pointed, at the same time, by two web pages (*A* and *D*) and, in order after it, there are those represented by the letters *D*, *A* and *B*. More details about the classification that has been applied here for them is below described.

Taking into consideration this test scenario, a demonstration in terms of iterations and obtained values for the web pages under examination is now shown, first through a scheme that has the purpose of explaining how each value was evaluated and then within a summary table through which the encountered results are illustrated. For convenience, to clearly explain this test case, each of the 4 web pages has assumed an initial PageRank value equal to 1, for which in the table below was not taken into account, and so transcribed, the so-called *iteration 0*. Also, the damping factor used in the example below described is a equal to 0.85, as suggested by the best test conducted by Sergey Brin and Lawrence Page about it, in order to get the most reliable results.

PR		**Equation to calculate PageRank**
PR(A)	$= (1-d) + d * \frac{PR(B)}{C(B)}$	$= 0,15 + 0,85 * PR(B)$
PR(B)	$= (1-d) + d * \frac{PR(A)}{C(A)}$	$= 0,15 + 0,425 * PR(A)$
PR(C)	$= (1-d) + d * (\frac{PR(A)}{C(A)} + \frac{PR(D)}{C(D)})$	$= 0,15 + 0,85 * PR(D) + 0.425 * PR(A)$
PR(D)	$= (1-d) + d * \frac{PR(C)}{C(C)}$	$= 0,15 + 0,85 * PR(C)$

Table 2.8: Assessment scheme adopted in the iterative algorithm of PageRank.

Iteration	Web page A	Web page B	Web page C	Web page D
N.1	1,00000000	0,57500000	1,42500000	1,00000000
N.2	0,63875000	0,57500000	1,42500000	1,36125000
N.3	0,63875000	0,42146875	1,57853125	1,36125000
N.4	0,50824844	0,42146875	1,57853125	1,49175156
N.5	0,50824844	0,36600559	1,63399441	1,49175156
N.6	0,46110475	0,36600559	1,63399441	1,53889525
N.7	0,46110475	0,34596952	1,65403048	1,53889525
N.8	0,44407409	0,34596952	1,65403048	1,55592591
N.9	0,44407409	0,33873149	1,66126851	1,55592591
N.10	0,43792177	0,33873149	1,66126851	1,56207823
N.11	0,43792177	0,33611675	1,66388325	1,56207823
N.12	0,43569924	0,33611675	1,66388325	1,56430076
N.13	0,43569924	0,33517218	1,66482782	1,56430076
N.14	0,43489635	0,33517218	1,66482782	1,56510365
N.15	0,43489635	0,33483095	1,66516905	1,56510365
N.16	0,43460631	0,33483095	1,66516905	1,56539369
N.17	0,43460631	0,33470768	1,66529232	1,56539369
N.18	0,43450153	0,33470768	1,66529232	1,56549847
N.19	0,43450153	0,33466315	1,66533685	1,56549847
N.20	0,43446368	0,33466315	1,66533685	1,56553632

Table 2.9: PageRank scores obtained applying the iterative algorithm on the test scenario.

To simplify the reading of the table it was chosen to limit to 8 the number of significant decimal digits, considering superfluous report further details about the obtained

PageRank values.

As we may notice, without entering into the merits of the goodness of the used algorithm and omitting its description as outside of the interest of this section, the application of the recursive calculation of PageRank involves the obtaining of values that seem to converge after only a few cycles of iterations, even if the analyzed case seems to be a scenario of drammatically reduced size compared to possible real cases.

In confirmation of what previously expressed, the experiment just conducted further demonstrates that PageRank is, as indicated above, a probability distribution between the different pages present on the web, being able to observe that in each case the sum of the obtained values is, in each cycle of iteration, equal to the sum of the initial PageRank scores of the web sites, or else equal to 4.

In any case, this example can easily be extended to more complex situations. Other investigations, in fact, were conducted, over time, by several people.

Other tests, for more complex situations than the one presented here, were shown and demonstrated subsequently by a new work by Sergey Brin and Lawrence Page in 1999 in [17], by which it is affirmed again the importance of this ranking algorithm.

Subsequently, a lot of other works have been done to better justify and understand the role of d, the damping factor, as shown in [5] and [14].

Another relevant work was done in [10] in which the authors tried to find and assess efficient techniques for computing PageRank, evaluating and studing the convergence properties, discussing the various aspects of its running time and, finally, demonstrating that PageRank can also be run on modestly equipped machines.

2.3 Temporal analysis

The world of the web, besides growing rapidly, is by nature a source of data whose mutation is an important and strong phenomenon, extremely present over time. Each information that populates it, in fact, can directly be updated by its author from time to time or, more simply, replaced by a new one, published on a new site or on a new web page.

So, the variety of the **temporal aspects** present, through different ways, on the web and related to the numerous news that colonize it, bring a very significant contribution, being useful for their better comprehension, especially at the level of awareness of their application context and their temporal location. Through these features, therefore, we are mainly able to verify the validity or not of a certain piece of information, establishing if during the period of interest and of reference it is still in force or no longer reliable. In 2015, therefore, to the question "President of the United States of America" a search engine that is based, among other things, on temporal information should return, between its results, only the documents (or the web pages) that refer to the name of Barack Obama, and no one else.

In general, it must be said, the research and the identification of temporal information can take place under various scenarios and points of view. Through NLP tools present at the state of the art, in fact, it is possibile to extract and capture them, regarding both the textual contents of documents, web pages, links that connect the latter and queries, and the meta-information that describe them, such as, for example, their creation date, their pubblication date, their update date and so on.

For the following treatments, to be clear, we refer to the first, or else to the temporal information observable inside their textual contents, as "Internal time" and to the second, or else to the time factor derivable by the meta-information, as "External time".

On the base of the analysis of these scenarios and these situations, a new area of the Information Retrieval, called **Temporal Information Retrieval (T-IR)**, began to

emerge. The hypothesis increasingly appreciated and supported in the context of the Web-IR, but not only, in fact, is to start taking into account the time factor and further "influence" the ranking of the web pages, for the purpose of contribute always more to the improvement of the quality of researches within large datasets.

It is not always easy to study and use information relating to the time. The systems that use them, in fact, should not simply shrink to the organization of the list of results on the basis of their "distance" from the involved temporal period.

Despite the great care taken with respect to this new problem, however, many of the techniques developed in this direction are not yet fully mature, so they can further be improved and used in new ways to increase the effectiveness of T-IR (and Web-IR) systems.

A first look at this fascinating area of research has been reported in [9], inside which the authors have tried to frame the major issues that will arise, immediately, following the massive growth of information resources available both on the web and not, hoping to powerfully support the searches. According to them, the IR (and T-IR) systems should continue to remain reliable and to provide acceptable solutions despite these difficulties. So, inside it, the article discussed some substantial differences between the classical Information Retrieval systems and those that include the time factor in their analyzes, stating that they should come across new and non-obvious challenges, such the extraction of temporal expressions and events, the temporal representation of documents, web pages, collections and queries, the realization of temporal retrieval models, the temporal and event-based summarisation, the temporal text similarity, the temporal query understanding, the clustering of search results by time, the temporality in ranking, the visualisation and design of temporal search interfaces, and so on.

Deepening these topics, another significative work was offered in [1]. According to the latter, the Information Retrieval applications do not take full advantage of all the temporal information embedded in documents to provide better search results and greater

user experience. In this work, so, it is presented a number of interesting applications along with open problems, in fact, despite what has been proposed and presented by the researchers at the state of the art, several challenging opportunities remain still unsolved. The goal is to discuss interesting areas and future work. They affirmed that T-IR systems must be considered as an extension of the existing ranking techniques, considering both the cases in which the temporal information are well-defined and the ones in which the temporal information must be organized to be useful. The authors have set themselves questions about the usefulness of the time factor to establish the similarity between two documents, or web pages, and the difficults to search for the temporal aspects within them. The treated problem, as said by the same authors, include several areas of computer science, mainly information retrieval, natural language processing, and user interfaces.

One notable study, to continue, is reported in [16]. It has explored two collections of web search queries to better investigate and understand the use of the temporal information needs. However, analyzing the time contained inside them, the authors found that, contrarily to their initial expectations, the use of temporal expressions in web queries is relatively scarce.

In [15], indeed, an interesting time-based language model is presented, based on TREC ad-hoc queries and documents. With this, the authors that proposed it tried to explain the relationship between time and relevance, considering the internal and external time of queries and the external time of documents, comparing these to some heuristic techniques.

A singular and important analysis, then, is shown in [8]. It is based on the concept that documents and queries are rich of temporal features. Studying and analyzing how use them, the goal that the authors tried to achieve is to exploit all internal and external temporal information to define new temporal scope similarities, proposing a new ranking model that combines the latter with the traditional keyword similarities.

They, so, have introduced the first non-probabilistic ranking model for the temporal scope of documents and queries in the world of the Temporal Information Retrieval, based on metric spaces.

Another important study is represented by [23]. Inside it, the main focus is addressed to the textual contents of documents. Starting from the temporal information that is possible to extrapolate from them, and enriching them with spatial information, the authors have presented a model for spatio-temporal document profiles, useful for many search and exploration tasks, like clustering, visualization or snippet generation.

A new interesting temporal model, name T-Rank and developed for the web, is shown in [3]. The authors tried to manage the several temporal information derivable from the creation, update, publication and removal time of web pages and links which connect them, with the introduction of a new concept called "freshness", thanks to which perform the ranking. The studies carried out about by them showed improvements in the quality of research on the web.

A different work, instead, is represent by [6]. Inside it, without primarily concentrating on the ranking and on the question of the ordering of the web pages and the quality of research, the authors studied the temporal information capturable from web page creation times and updating times, to better explain the dynamicity that characterize the web.

Further analysis on the temporal information in the reality of the web, is offered in [2]. In the latter, in fact, is highlighted the importance of the external time present in the links, showing that if the search engines would be able to track e save them, then, they would be able to provide results that are more timely and better reflect current real-life trends than those they provide today.

Finally, a consideration must be made. The classical web ranking algorithms often miss the temporal dimension, an important dimension. The web, in fact, is not a static environment, but it changes constantly. In [25] and [24], this heavy failings are

discussed and deepened. Consequently to these hypotheses, the quality of some pages in the past may not be the same now or in the future. The ranking algorithms, hence, should consider the external time factor, analyzing it through the present links, trying to understand how it contributes to the potential importance of a web page and how it and its influence change over time. In these works, so, are presented two new algorithms named, respectively, TimedPageRank and TemporalRank, corresponding to two modified versions of the original PageRank method.

To conclude, finally, a summary table concerning the time factor and some of the solutions currently known to the state of the art is below presented, in order to show how and where this significant phenomenon is treated within documents, web pages, links, queries and even more. In analyzing this, moreover, the table highlights, inside it, the shortcomings that have not yet been taken into account and remedied by the scientific community.

	Internal time	**External time**
Documents	[8], [23]	[15], [8]
Web pages		[3], [6]
Links		[25], [24], [3], [2]
Queries	[16], [15], [8]	[15], [8]

Table 2.10: Summary of the time factor analysis in known solutions.

2.3.1 Temporal expressions

The time, as already widely known, is the subject of several studies in many disciplines and it plays a decidedly key role with regard to the understanding of a web page or a

document, being able to correctly describe its chronological and application context. The temporal aspects, of course, can be expressed and manifested in different ways, both in relation to the web pages and within the same queries.
The modes in question, so, are the following:

- **Explicit**: it consists of a direct reference to a specific moment in time and this does not require additional knowledge to allow its interpretation. An example of this is the phrase *"March 2016"*.

- **Implicit**: it consists of an indirect reference to a specific moment in time and this requires additional knowledge to allow its interpretation. An example of this consists of the names of holidays, events, or anything else known as *"Christmas 2015"*.

- **Relative**: it consists of an indirect reference to a specific moment in time and this requires the knowledge of the context in which it is expressed. An example of this is the word *"Today"*.

In the figure below, to better explain what just said, it is represented a portion of a sample web page, named *Story.html* and created by its author in the 1998-04-18, in which there are some temporal references that can be exploited to understand its chronological context, and so its application context and domain.

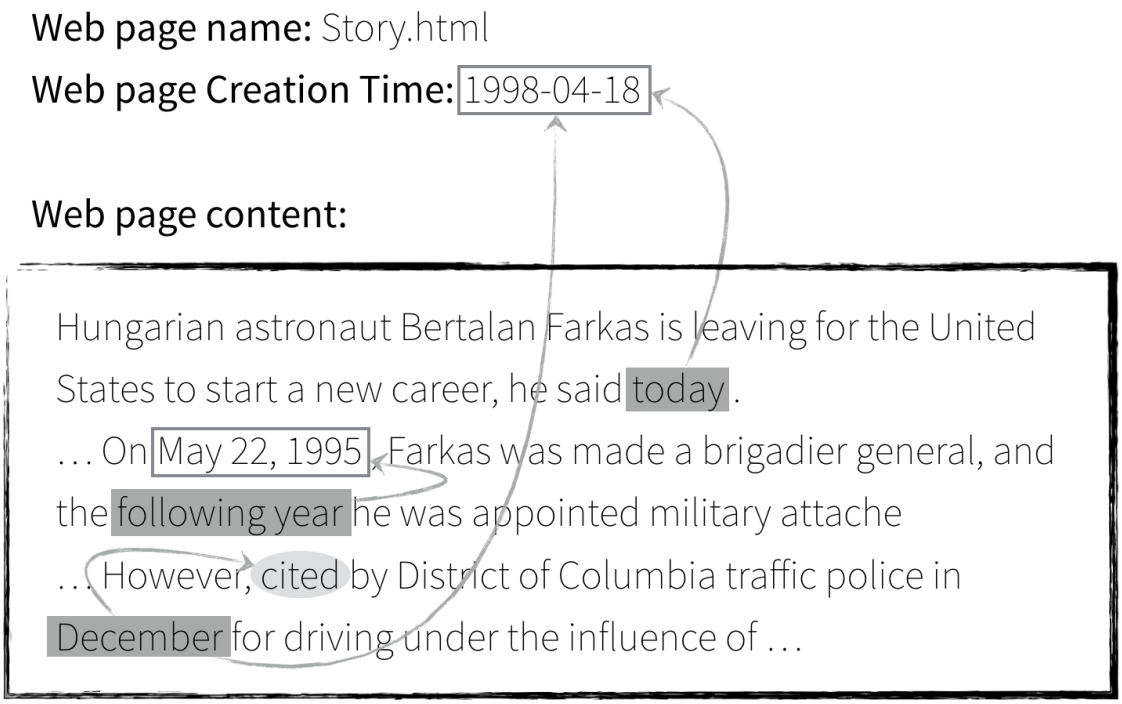


Figure 2.2: Example of some temporal aspects present in a sample web page.

Based on what has been previously achieved, so, it is possible to point out and introduce some terms and definitions in this regard, useful for allowing a better interpretation of the temporal aspects, such as:

- **Chronon**: it corresponds to the (atomic) unit of measurement used for the time and its size is chosen according to the dataset in which there is the need to analyze and inspect it.

- **Timeline**: it is equivalent to the entire set of all possible temporal information, in chronological order and, so, it constitutes the real and effective time domain,

in which the events happen.

- **Temporal expression**: it corresponds to a temporal interval or a temporal period, comprising pairs of order chronons, belonging to the time domain (timeline), indicating, as a result of this, a time duration.

2.3.2 Management of temporal expressions

In the previous sections, the art and the importance of the temporal analysis was introduced, mainly focusing on the reasons for which this has to be conducted and the benefits that can be drawn through its use, specifying, then, where it is possible to locate and identify the time factor.
Subsequently, therefore, the temporal expressions that can be obtained and exploited and their composition have been described, in order to better understand their chronological and application contexts, which can make more or less relevant a web page or, more generally, a simple document with respect to the currently considered reference query.
At the state of the art, however, there are several methods and numerous solutions that allow and permit the comparison between two different temporal expressions, by which, therefore, properly manage their meaning and the contents indicated inside them.
Before analyzing how this task was addressed and completed, however, it is necessary to present some key concepts that allow a better interpretation and, above all, a correct comparison between temporal expressions.
These are:

- **Overlap**, with this term it is meant that two temporal expressions overlap each other, or else they share a period in common. A simple case of overlap, for example, is present between the following temporal expressions:

    1. Start: *"15 March 2016"*.
       End: *"15 April 2016"*.

    2. Start: *"01 March 2016"*.
       End: *"31 March 2016"*.

- **After**, with this term it is meant that a temporal expression follows another one. In this sense, indeed, it is enought think to the two following temporal expressions:

    1. Start: *"01 April 2016"*.
       End: *"31 April 2016"*.

    2. Start: *"01 March 2016"*.
       End: *"31 March 2016"*.

- **Before**, with this term it is meant that a temporal expression precedes another one. An example that covers this case is the exact opposite of the previous treated:

1. Start: *"01 April 2016"*.
   End: *"31 April 2016"*.

2. Start: *"01 March 2016"*.
   End: *"31 March 2016"*.

Among the different existing and recognized (by the scientific community) methods, so, it is worth mentioning someone.
Considering now the already defined test scenario and focusing on the web page P and the query Q, each containing a temporal expression, respectively named T_P and T_Q, there are:

- **Manatthan distance**, or Taxicab geometry, it considers the distance between two points as the sum of the absolute differences between the starting and the ending times of two temporal intervals, resulting zero only when they are exactly the same. It is a symmetric method, since it sums up the distance in absolute value, without knowing if an interval is starting from a query or a web page:

$$MNH(Q, P) = |T_Q.Start - T_P.Start| + |T_Q.End - T_P.End|$$

- **Query-biased hemidistance**, it takes into account the possibility in which a user might consider the broader web page more relevant because it covers the query temporal period, assigning a distance equal to zero to all web pages that completely cover it, and a positive distance to the ones that do not cover part

of it. Furthermore, if the query and temporal periods of the web page and the query do not intersect, the gap between them is also added to their distance. In contrast to the previous method, this is not symmetric:

$$QBH(Q, P) = (T_Q.End - T_Q.Start)$$

$$-(min\{T_Q.End, T_P.End\} - max\{T_Q.Start, T_P.Start\})$$

- **Document-biased hemidistance**, in the opposite case with rispect to the previous one, it takes into account the possibility in which a user might consider the broader web page less relevant because it is too generic, evaluating more relevant the one that falls inside the temporal interval represent by the query:

$$DBH(Q, P) = (T_P.End - T_P.Start)$$

$$-(min\{T_Q.End, T_P.End\} - max\{T_Q.Start, T_P.Start\})$$

in which:

- $\boldsymbol{T_Q.Start}$ and $\boldsymbol{T_P.Start}$: they represent the start times of the temporal interval represented by Q e P.

- $\boldsymbol{T_Q.End}$ and $\boldsymbol{T_P.End}$: they represent the end times of the temporal interval represented by Q e P.

- $\boldsymbol{max\{T_Q.Start, T_P.Start\}}$ and $\boldsymbol{max\{T_Q.End, T_P.End\}}$: the function returns, as output, the maximum value between the two considered parameters.

- $\boldsymbol{min\{T_Q.Start, T_P.Start\}}$ and $\boldsymbol{min\{T_Q.End, T_P.End\}}$: the function returns, as output, the minimum value between the two considered parameters.

To better understand the role of the methods described above, an illustrative figure representing an example of their use is below reported, treating five possible different cases between a web page P and a query Q. In the latter illustration, by detailing, will be shown the scenarios in which Q and P present the same temporal expression, completely different temporal expressions, a temporal expression that entirely covers the other, and a temporal expression that intersect between them.

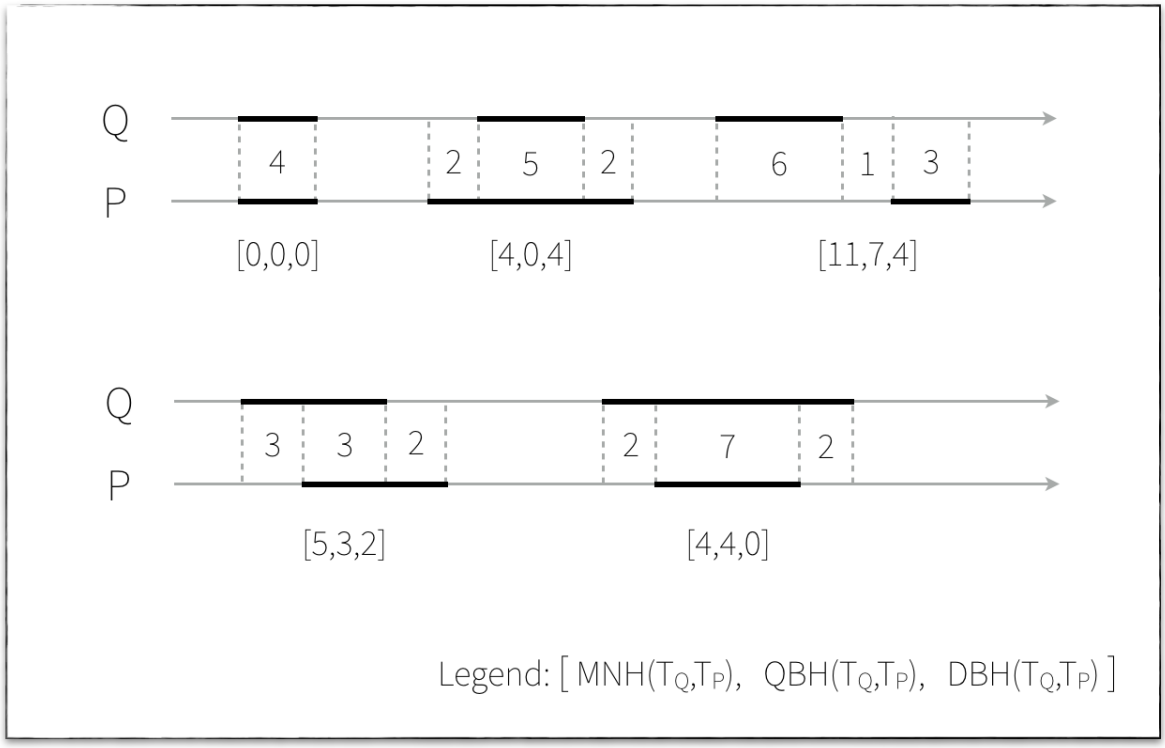


Figure 2.3: Example of management of temporal expressions through methods known to the state of the art.

In the used depiction, all the comparison between temporal expressions are accompanied by a set of three values. In the latter, the first corresponds to the result reached by the *Manatthan* method, the second to that obtained by the *Query-biased hemidistance* method and the last, finally, that encountered by the *Document-biased hemidistance* method.

Chapter 3

Temporal PageRank

"Getting information off the Internet is like taking a drink from a fire hydrant."

— Mitchell Kapor

This thesis work, as anticipated, aims to improve the state of the art with regard to Web-IR systems, providing a contribution which is based on the current existing proposals, on the many ideas presented and discussed by others about, on the possible applications, on open problems and challenges that can differently be overcome or dealt, in order to be able to better satisfy the informative needs of users.
The motivational conditions, so, are a lot, for which starting from those a new measure has been defined for judging the relevance of the web pages retrieved on the web, in accordance with the query that interrogated the search engine.
The expressed concepts, consequently, have allowed the definition of *Temporal PageRank (T-PR)*, a new temporal model for the Web Information Retrieval, provided by leveraging and combining between them different factors, so as to judge each item retrieved by the search engine on the basis of, in the first place, the importance of the authors (or else of the sources) who created the information described inside the web pages and, in a second time, the temporal aspects that make more or less reliable the

data contained within them. They are, respectively:

1. **PageRank**, as discussed before, the famous ranking algorithm by which analyze and understand the relevance of the web pages that contain the desired information, being able to synthesize an objective measure of their importance starting from the subjective ideas of users about, suggesting, so, how rank them on the base of their sources.

2. **New temporal model**, created with the aim to capture the temporal aspects present, at the same time, both in queries and in web pages, so as to ascertain which of these latter best satisfy the (temporal) informative needs expressed by the users. The depth description of this new model is discussed in the following chapter.

The goal, combining in the right way the components presented here, is to give greater significance to those information that are compatible with the expressed temporal desire present, in different forms, in the queries that interrogated the web, and not more just on the only base of the simple textual similarity that they reached, simultaneously giving a greater relevance to those news coming from all those that, publishing them, have at that time a better importance, or else all those that are considered the most reliable sources.
Before the final formulation of *Temporal PageRank*, however, have been examined various hypotheses, according to which, after evaluating them in detail, some first different temporal models were then processed. This temporal models take into account the time factor within the links that connect two or more web pages in the collection, inside the

recovered or to be recovered web pages themselves, in the queries used to interrogate the web and even more.
Finally, an additional and complete temporal model, based on these latter, has been proposed. This considers and covers them, assessing, so, the possible contemporary presence of temporal aspects within links, web pages and queries.
In this chapter, therefore, the formalization of *Temporal PageRank* is fully discussed, describing in the first place the composition of each factor taken into account and, in detail, the temporal models designed for it. These latter, in fact, are evaluated, studying how they have impact on the proposed Web Information Retrieval system, and how the most promising between them is then chosen and combined with the other factors, first presented, under analysis.

3.1 Temporal model

The design and the development of a temporal model, as expressed, have been undertaken with the intention of studying and analyzing the temporal needs expressed by users during their searches, in order to allow, therefore, to the Web Information Retrieval system, to privilege the recovery of those web pages that appear temporally closer to them and which seem to temporally satisfy them in a better way.
In the state of the art, however, there are some already in existence solutions that consider the temporal aspects contained within the links and that allow, consequently, the PageRank method to assess the entity of the impact and the influence that these have regarding the sorting of those web pages deemed relevant or not, as anticipated and discussed in the previous chapters.
At the same time, in contrast, a proposal that is able to study and understand the time factor contained within the textual contents that compose the web pages present inside

the collection under examination, or in general that constitute the web, is lacking in the scientific community.
This section, accordingly, has the duty to treat the in-depth description of the new developed temporal models, which are, then, adopted during the implementation of the proposed Web Information Retrieval system.
Once the temporal expressions are identified inside the collection under analysis, these were handled and compared between them through the *document-biased hemidistance* method, first introduced and presented, already known to the scientific community.

3.1.1 Considering (external) time in links

A first temporal model has been developed to provide a privileged importance to the external time contained in the links present between the various pages belonging to the world of the web, so as to consider as the most important the incoming contributions coming from the most recent links pointing to some certain web pages.
The idea, in considering this, is that the best solutions, in this case the best web pages, are those that appear to be continuously fed in time. In this way, in fact, the intuition of assess, as the most relevant, the derived contributions coming from the most recent links was born, considering them the best ones.
To do this, therefore, it was necessary to compare and judge the temporal expressions described by the links with the current time in which there is the need to study the used collection of data, so as to decrease their importance as much as they are old.
More specifically, so, given a generic web page P, let *Curr_time* be the current time, let *DBH* be the document-biased hemistance method to manage the temporal expressions and on the basis of the original definition of PageRank and the time contained in the links, the new score is defined as:

$$PR_{link-time}(P) = (1-d)+$$

$$+d(\frac{PR(P_1)}{(DBH(Curr\_time, P_1)+1) * C(P_1)} + \ldots + \frac{(PR(P_n)}{DBH(Curr\_time, P_n)+1) * C(P_n)})$$

The model, therefore, proves to be independent with respect to the queries submitted by the user to the interested search engine, which in fact, does not consider them. To better understand what has just reported, an example relative to the test scenario is below discussed. To do that, however, the latter was enriched with additional information through which were announced all the temporal expressions (each specified by a beginning and an ending time) expressed from the considered web pages (considering every one observable inside them) and the links (in which the start and the end time corresponds to their creation date) that connect them.

Below, so, two figures are shown.

A first one is related to the enriched test scenario, in which the relevance of the web pages is given exclusively by the calculation of their original PageRank score and a second one that considers the (external) time factor in the links, thus showing the impact that this solution has on the importance of the elements which are part of the analyzed web collection of data.

In the first picture, to be precise, inside each of the web pages, represented by a circle, the temporal expressions that can be derived through their contents are described, while near each of the links belonging to the scenario, each green box contains the temporal expressions relating to it.

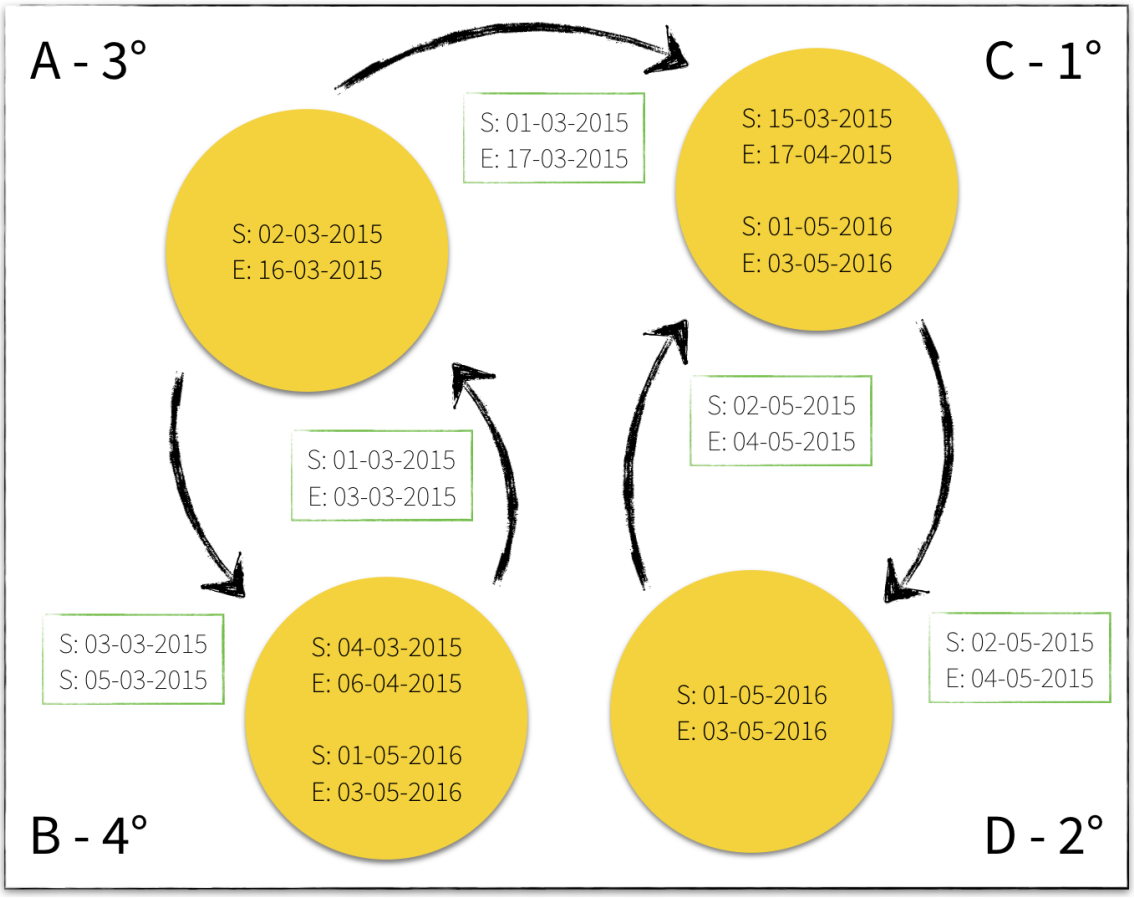


Figure 3.1: Test scenario enriched with temporal expressions. Inside the circles there are those related to the web pages content, in *green* those concerning the links. In the corners are indicated, for each page, the positions reached in the ranking.

For the other, instead, remembering that the temporal aspect is not considered within the query, the reference is made to the current time relative to the execution performed on the test scenario, hypothesizing for simplicity, that it is equal to *Curr_time*, such that $T_Curr\_time.Start$ is equal to 04/05/2015 and $T_Curr\_time.End$ corresponds to 04/05/2015.

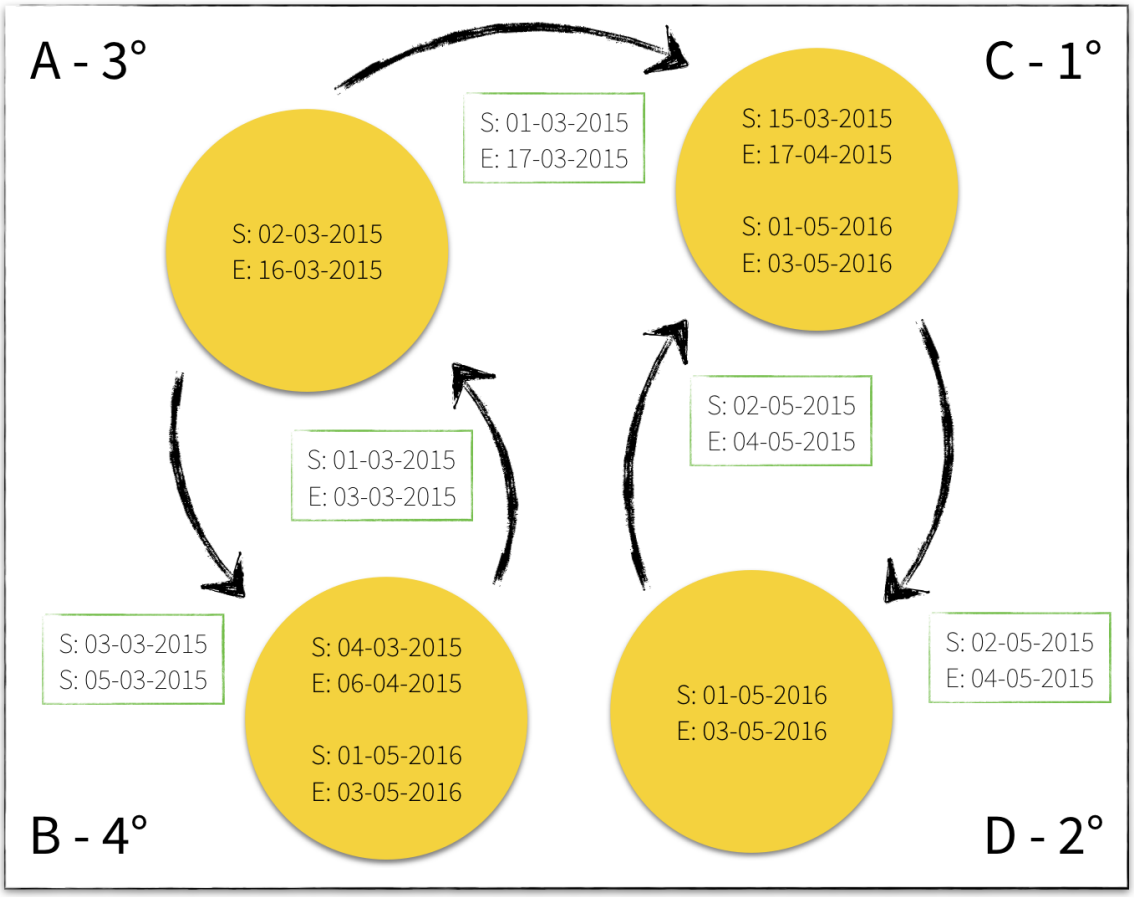


Figure 3.2: Importance of web pages in test scenario considering time in links. Inside the circles there are those related to the web pages content, in *green* those concerning the links. In the corners are indicated, for each page, the positions reached in the ranking.

Link	Document-biased Hemidistance
A → B	62
A → C	64
B → A	64
C → D	2
D → C	2
Original PageRank ranking	C, D, A, B
$PR_{link-time}$ ranking	C, D, A, B

Table 3.1: Updated PageRank ranking considering time in links.

The used example, however, does not upset the rankings established by the original version of the PageRank method, confirming all the placings that the considered web pages had managed to conquer.
This because the only links that demonstrate to have temporal expressions useful for the purpose to be achieved are those present between C and D.
Nevertheless, even without evident changes, the temporal model which analyzes the time factor within the links present between the various pages, provides an additional contribution that allows to emphasize, even more, the importance of the pages C and D, more clearly insisting their superiority over the other pages A and B.

3.1.2 Considering (external) time in links and (internal) time in queries

A second temporal model was then developed in order to study and exploit the external time factor existing in the links that connect two or more web pages between them

and the internal time manifested by the queries submitted to the search engine.
As originally expressed by the PageRank method, a web page assumes a greater importance with the increase of the owned incoming links or upon the occurrence of the growth of those considered important (they are so in the cases where these coming from sources deemed authoritative).
The idea, therefore, is to assign a weight to each of the contributions bring by incoming links, in such a way as to reward those temporally closer and more pertinent with respect to the temporal period made known by the used query.
The calculation of the new PageRank score, in accordance with the time contained and exposed by the links that connect the pages belonging to the web collection and the reference query, gives a new ordering to the obtained results, such that, let DBH be the document-biased hemistance method to manage the temporal expressions, for a generic web page P and a query Q, is defined as:

$$PR_{link+query-time}(Q, P) = (1 - d)+$$

$$+d(\frac{PR(P_1)}{(DBH(Q, P_1) + 1) * C(P_1)} + \ldots + \frac{PR(P_n)}{(DBH(Q, P_n) + 1) * C(P_n)})$$

To better understand what has just reported, an example relative to the test scenario, presented in the Figure 3.1 and enriched with temporal expressions, is below shown, considering the query Q, such that $T_Q.Start$ is equal to 03-03-2015 and $T_Q.End$ corresponds to 05-03-2015, in order to evaluate the new importante gained by the web pages.

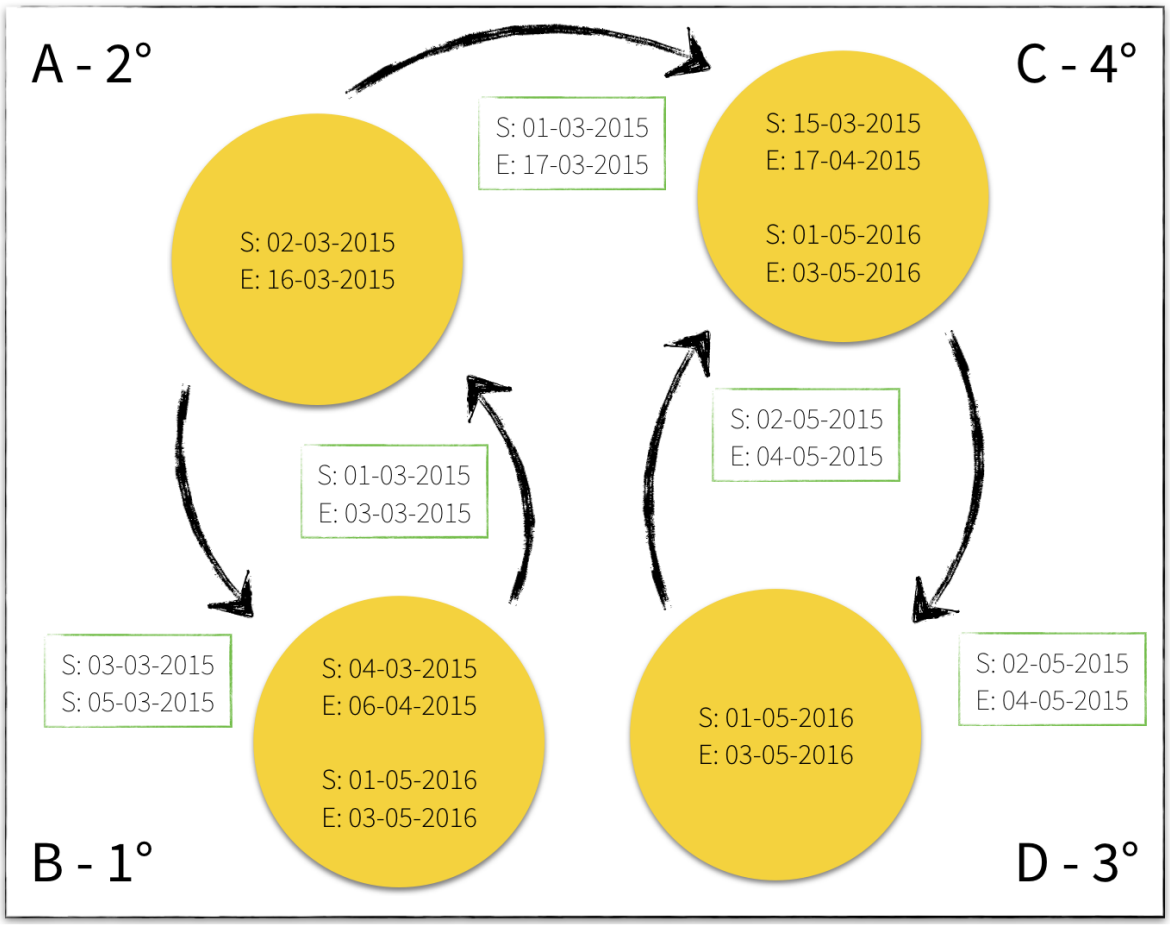


Figure 3.3: Importance of web pages in test scenario considering time in links and queries. Inside the circles there are those related to the web pages content, in *green* those concerning the links. In the corners are indicated, for each page, the positions reached in the ranking.

Link	Document-biased Hemidistance
A $\rightarrow$ **B**	0
A $\rightarrow$ **C**	14
B $\rightarrow$ **A**	2
C $\rightarrow$ **D**	58
D $\rightarrow$ **C**	58
Original PageRank ranking	C, D, A, B
$PR_{link+query-time}$ ranking	B, A, D, C

Table 3.2: Updated PageRank ranking considering time in links and queries.

Analyzing this example, it is clear that the links between the web pages C and D represent the worst incoming contribute, due to not pertinent temporal expressions with respect to the one expressed by Q, the used query. The best links, instead, are those that connect the web pages A and B, because both they cover and satisfy the temporal informative need. As a consequence of this, so, the web pages contained and described inside the test scenario gain new importance scores, giving life to a new ranking. In the latter, in fact, the web page B go up from the last position to the first one, togheter with the web page A, while C and D are relegated to the last placements, at the bottom, so, of this new classification.

3.1.3 Considering (internal) time in web pages and (internal) time in queries

A third temporal model was then developed in order to understand and assess the temporal impact of the temporal expressions observable and present inside the web

pages included in the collection under review and within the queries used by users to express their (temporal) informative needs.
Considering that this model does not consider the links, but it evaluates only the internal contents of the web pages, it can easily be applied to a suite of simple documents and not necessarily to a web collection.
The model, so, is opportunely developed with the aim of giving a score as inferior to those web pages whose content is more distant and less pertinent, from a temporal point of view, by the temporal need represent by the considered query.
The new method, designed on the base of the original definition of PageRank and on the time contained and described inside the web pages and the queries, provides a new ordering to the obtained results through the calculation of a new score, for a generic web page P and a query Q, such that, let DBH be the document-biased hemistance method to manage the temporal expressions, the importance of P decreases with respect to its temporal expressions, representing, as already said, its chronological contexts, as shown below:

$$PR_{content+query-time}(Q, P) = \frac{PR(P)}{DBH(Q, P) + 1}$$

As with the previous models, a useful example is below reported to simplify the understanding of what has just pointed out, accompanying the latter with a summary table.
Subjecting the test scenario (Figure 3.1) to a query Q, such that it describes, inside it, a temporal expression in which $T_Q.Start$ is equal to 02-03-2015 and $T_Q.End$ corresponds to 16-03-2015, in fact, each of its web pages gains or loses part of its relevance, in relation to the temporal content that they represent.

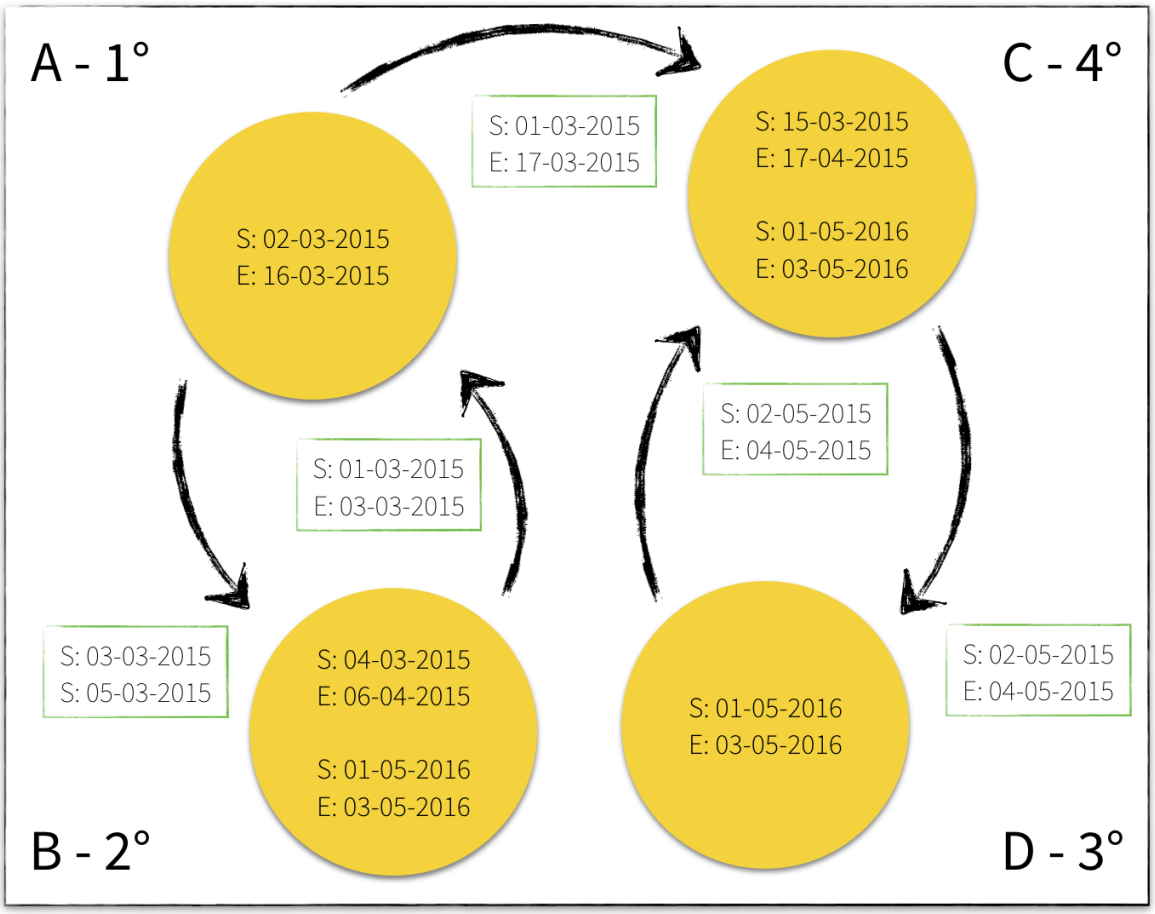


Figure 3.4: Importance of web pages in test scenario considering time in web pages and queries. Inside the circles there are those related to the web pages content, in *green* those concerning the links. In the corners are indicated, for each page, the positions reached in the ranking.

Web page	Document-biased Hemidistance
A	0
B	447
C	447
D	412
Original PageRank ranking	C, D, A, B
$PR_{content+query-time}$ ranking	A, B, D, C

Table 3.3: Updated PageRank ranking considering time in web pages and queries.

As evidenced by this reproduction, therefore, the ranking is completly changed. The more pertinent from a temporal point of view is the web page A, with a document-biased hemidistance value equal to 0, considering that it possess the same temporal expression expressed by the query. After it, also the web page B goes up in the classification, gaining the second placement. Become, instead, the less relevant web pages, the ones represent by the letters D and C.

3.1.4 Considering (external) time in links, (internal) time in web pages and (internal) time in queries

A last temporal model, finally, was developed to introduce a new modified version of the PageRank method, conceived on the basis of the previously presented ones, to allow the taking into account of the time factor existing, at the same time, within web pages, links that connect them in the world of the web and used queries. In this sense, then, given a generic web page P, a query Q and let DBH be the document-biased hemistance method to manage the temporal expressions, this new temporal variant

has been then defined as shown below:

$$PR_{link+content+query-time}(Q,P) = \frac{PR_{link+query-time}(Q,P)}{DBH(Q,P)+1}$$

This model, as can be imagined, decreases the importance of a web page, calculated on the basis of the external time factor of links and the temporal expressions described inside the web pages content, with respect to the temporal context that is represented by the reference query, considering, then, the time factor under many points of view.
To clarify what has been developed and highlighted above, a representative figure is below reported, including a possible case of application of this PageRank variant on the test scenario, annunced in Figure 3.1.
To do this, let Q be the considered query, such that $T_Q.Start$ corresponds to 03-03-2015 and $T_Q.End$ is equal to 05-03-2015.

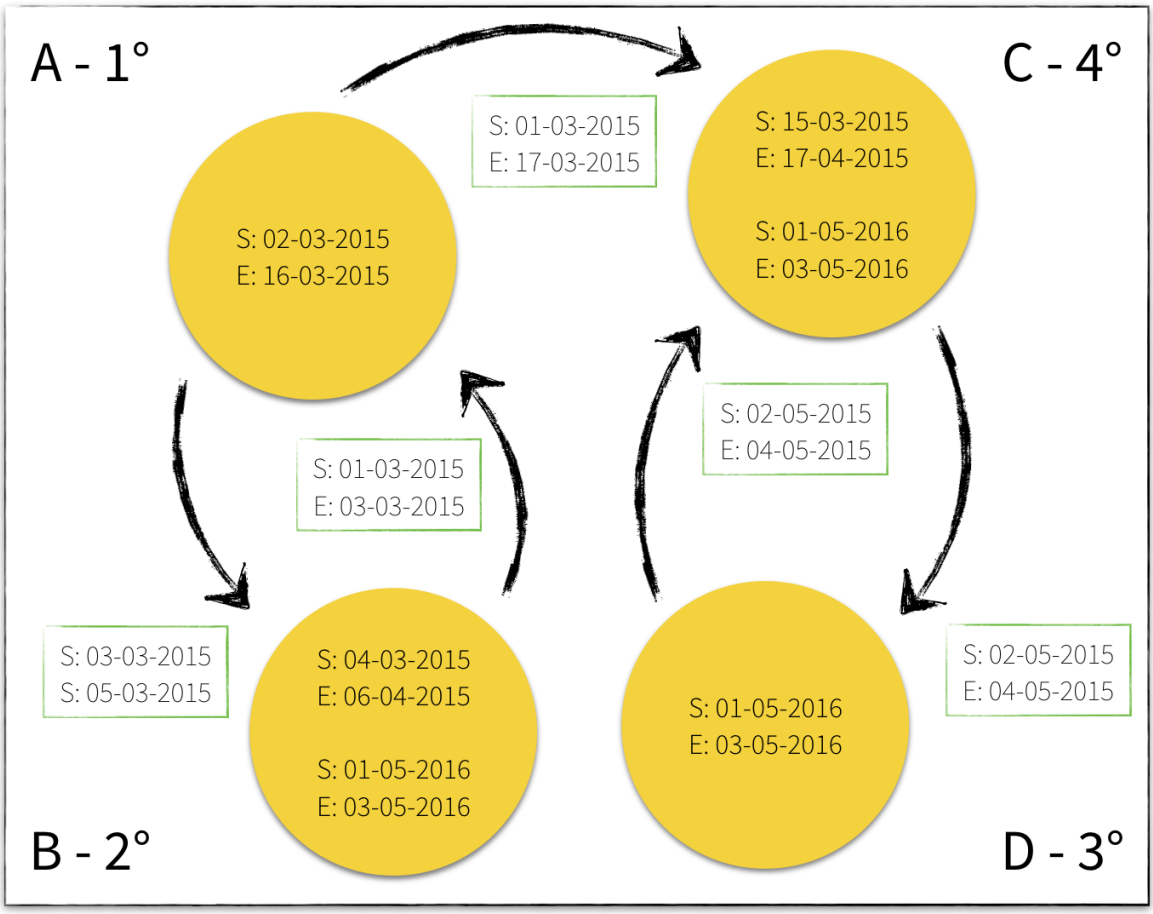


Figure 3.5: Importance of web pages in test scenario considering time in links, web pages and queries. Inside the circles there are those related to the web pages content, in *green* those concerning the links. In the corners are indicated, for each page, the positions reached in the ranking.

Element	Document-biased Hemidistance
Link **A** $\rightarrow$ **B**	0
Link **A** $\rightarrow$ **C**	14
Link **B** $\rightarrow$ **A**	2
Link **C** $\rightarrow$ **D**	58
Link **D** $\rightarrow$ **C**	58
Web page **A**	12
Web page **B**	458
Web page **C**	458
Web page **D**	425
Original PageRank ranking	C, D, A, B
$PR_{link+content+query-time}$ ranking	A, B, D, C

Table 3.4: Updated PageRank ranking considering time in links, web pages and queries.

From what can be seen, therefore, also in this case, it is extremely interesting that the list initially drawn up by the original PageRank method is almost completely changed. The web pages B and A, in fact, rise to the top and become the most relevant and important in the test scenario, best representing the temporal informative needs. The web pages C and D, instead, conclude in the last positions of this new ranking, as a result of not useful temporal expressions contained both in links and in their textual content.

3.2 Combining the factors

After defining the main factors that are behind the development of *Temporal PageRank*, the modes adopted to combine each other are presented in this section, illustrating, finally, as they give life to the complete formulation of *T-PR*.
To conduct the tests that have been specially built and developed to challenge it, it should be specified, several its variants were initially considered, as already anticipated, in order to be able to choose and adopt, subsequently, the one which, between them, shows the ability to achieve and gain the most significant and important results, with respect to the collection of web pages under analysis.
Before proceeding, however, a summary of the features of the new developed temporal versions of the PageRank method is reported in the table below.

	PageRank	**Links time**	**Web pages time**	**Queries time**
$PageRank$	X			
$PR_{link-time}$	X	X		
$PR_{link+query-time}$	X	X		X
$PR_{content+query-time}$	X		X	X
$PR_{link+content+query-time}$	X	X	X	X

Table 3.5: Summary of the developed PageRank variants.

After that, on the basis of the new created PageRank variants and combining in a suitable way the involved factors between them, five different interpretations of *Temporal PageRank* have been developed, considering, in addiction, the textual similarity reached by the web pages with respect to the textual content expressed by the used queries:

- $\boldsymbol{PageRank_{textual}}$, it is based on the original formulation of the PageRank method and does not consider the time factor:

$$PageRank_{textual}(Q, P) =$$

$$= (1 - \alpha)BM25(Q, P) + \alpha * PR(P)$$

- $\boldsymbol{TPR_{link-time}}$, it considers the (external) temporal factor of the incoming and outgoing links of the web pages:

$$TPR_{link-time}(Q, P) =$$

$$= (1 - \alpha)BM25(Q, P) + \alpha * PR_{link-time}(P)$$

- $\boldsymbol{TPR_{link+query-time}}$, it considers the (external) temporal factor of the incoming and outgoing links of the web pages and the (internal) time of the reference queries:

$$TPR_{link+query-time}(Q, P) =$$

$$= (1 - \alpha)BM25(Q, P) + \alpha * PR_{link+query-time}(Q, P)$$

- $\boldsymbol{TPR_{content+query-time}}$, it considers the (internal) temporal factor within the textual contents of the web pages and the reference queries:

$$TPR_{content+query-time}(Q, P) =$$

$$= (1 - \alpha)BM25(Q, P) + \alpha * PR_{content+query-time}(Q, P)$$

- $\boldsymbol{TPR_{link+content+query-time}}$; it considers the (external) temporal factor of the incoming and outgoing links of the web pages and the (internal) time of their textual contents and of the reference queries:

$$TPR_{link+content+query-time}(Q, P) =$$

$$= (1 - \alpha)BM25(Q, P) + \alpha * PR_{link+content+query-time}(Q, P)$$

in which:

- $\boldsymbol{\alpha}$: it is the component responsible for regulating the linear combination between the *BM25* factor, which handles the textual similarity between question and answer and the modified versions of the *PageRank* method, among the different

considered. Its value, it must be specified, has been chosen only in a second moment, after having carried out the relative tests, in order to be able to adopt the one that, in accordance with the obtained results, confirmed to be capable of ensuring the more functional configuration. In any case, therefore, this can appropriately be set according to the collection of data that is currently under examination.

In any case, it must be said, every *Temporal PageRank* interpretation appears to be, so, dependent with respect to the queries submitted to the search engine.

Chapter 4

Experimental evaluation

"If we knew what we are doing, this not could be called research, right?"

— Albert Einstein

After formulating *Temporal PageRank*, according to what has just been discussed in the previous chapters, it is necessary to deepen and examine in detail its properties and its most significant features, testing and comparing it with the current solutions known to the state of the art, in order to find out if the latter proposal is able to ensure the possibility to achieve superior and preferable results and result, consequently, more useful, in the context of the Web Information Retrieval, to the scientific community and not only.

To rightly do this and in the most effective way, there is the need to properly define and configure the environment in which it is evaluated and judged, in order to, then, figure out what are its best features and what are, instead, those that are less convincing on which we must still work.

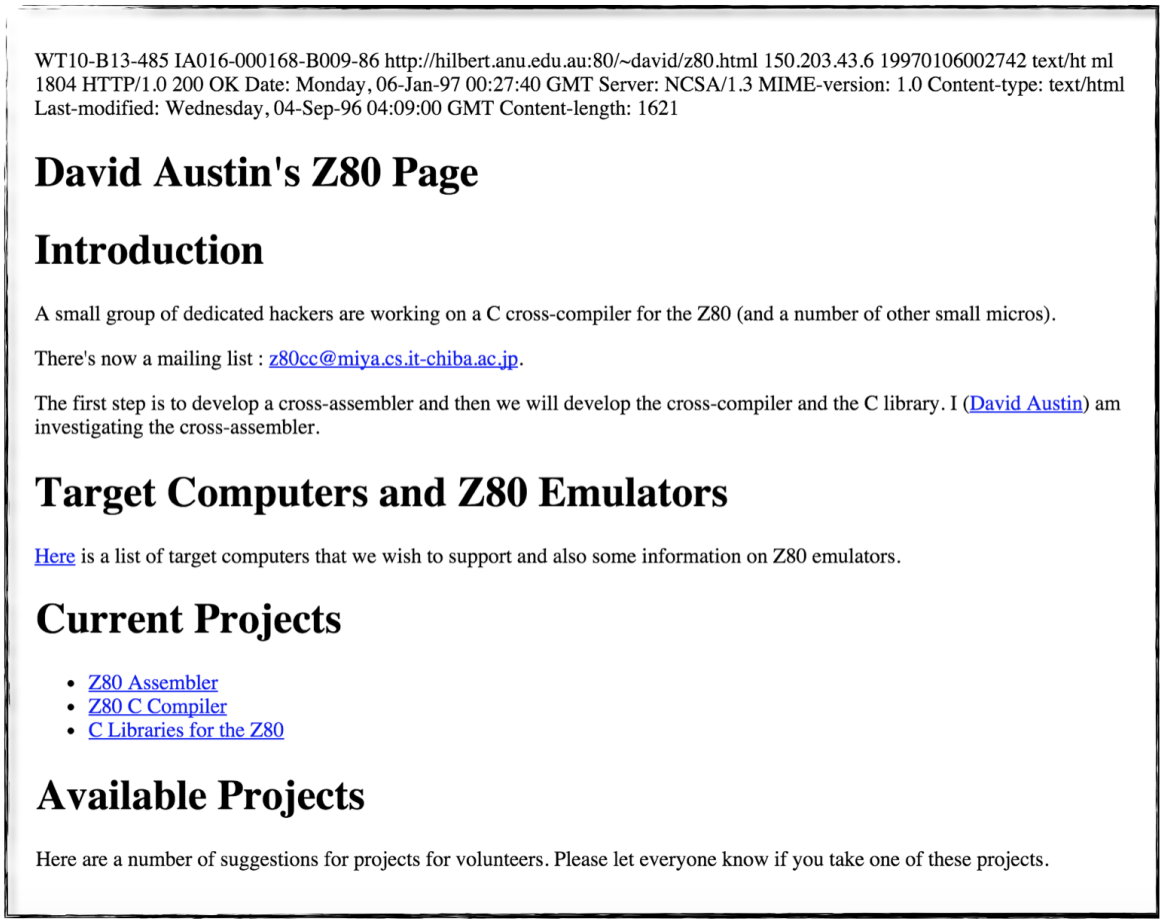
In the discussion of this chapter, first of all, the reference dataset used to perform the searches and test *Temporal PageRank* is presented, describing how this is constituted and all its most important behaviors, such as types and formats of the data contained

inside it, languages used for them and much more, by providing, if necessary, the representation of an example file.
Subsequently, to continue, all the evaluation measures taken to study every achieved result are established, defining, in addition, how the realized model is compared with other known solutions, to see if, indeed, it is able to make improvements with respect to them.
Before getting to the results found, however, a goal of this chapter is to explain and discuss, in detail, the contribute that each factor present in *Temporal PageRank* is able to bring to the system, analyzing how they are used to perform the tests on the considered reference dataset.

4.1 Dataset

The collection of data adopted to achieve the prefixed goal, so, is the *Web TREC WT2G* [1] collection, widely known in the field of the Information Retrieval and not only, presented and discussed in [12], and it consists of a large set by the total size of 2GB, including 247491 web pages.
The details of its composition are shown below, first through a figure representing a sample web page [2] present inside the collection in question and then thanks to a summary table containing its most important and significative features.

[1] Accessible through `http://ir.dcs.gla.ac.uk/test_collections/`
[2] Accessible via `http://ir.dcs.gla.ac.uk/test_collections/samples/wt2g_sampleDoc`



WT10-B13-485 IA016-000168-B009-86 http://hilbert.anu.edu.au:80/~david/z80.html 150.203.43.6 19970106002742 text/ht ml 1804 HTTP/1.0 200 OK Date: Monday, 06-Jan-97 00:27:40 GMT Server: NCSA/1.3 MIME-version: 1.0 Content-type: text/html Last-modified: Wednesday, 04-Sep-96 04:09:00 GMT Content-length: 1621

David Austin's Z80 Page

Introduction

A small group of dedicated hackers are working on a C cross-compiler for the Z80 (and a number of other small micros).

There's now a mailing list : z80cc@miya.cs.it-chiba.ac.jp.

The first step is to develop a cross-assembler and then we will develop the cross-compiler and the C library. I (David Austin) am investigating the cross-assembler.

Target Computers and Z80 Emulators

Here is a list of target computers that we wish to support and also some information on Z80 emulators.

Current Projects

- Z80 Assembler
- Z80 C Compiler
- C Libraries for the Z80

Available Projects

Here are a number of suggestions for projects for volunteers. Please let everyone know if you take one of these projects.

Figure 4.1: Example of a WT2G web page.

WT2G	**Information**
Web pages	247491 web pages (for 2GB of total)
Relevant web pages	2279/247491 (0,92% density)
Links	1166702 links (about 4.71 links for web page)

Table 4.1: WT2G collection features.

Inside it, however, beyond the data regarding the composition itself which characterizes every single present web page, there are important and interesting additional information, for each of them, used for conducting the analysis and for the initial configuration of *Temporal PageRank* that had permission to properly execute it and validate, then, its reported results.

Among the main peculiarities that can be derived from the information specified above there is, without doubt, the structure that identifies and represents the network. The latter, in fact, can easily be managed by reading the *inlinks.txt* file, in which all the links that connect two or more web pages are specified, as below shown through an its sample part:

Web page's ID	**Set of web pages containing links direct to the first**
WT01-B01-1	WT01-B01-2 WT01-B01-3 WT01-B01-4 WT01-B01-5
WT01-B01-10	WT01-B01-7 WT01-B01-8 WT01-B01-9
	WT01-B01-11 WT01-B01-15
WT01-B01-100	WT01-B01-42
WT01-B01-101	WT01-B01-42
WT01-B01-102	WT01-B01-42
...	...
WT28-B01-95	WT28-B01-26 WT28-B01-96
WT28-B01-96	WT28-B01-26 WT28-B01-95 WT28-B01-97
WT28-B01-97	WT28-B01-26 WT28-B01-96 WT28-B01-98
WT28-B01-98	WT28-B01-26 WT28-B01-95 WT28-B01-96
	WT28-B01-97 WT28-B01-99
WT28-B01-99	WT28-B01-26 WT28-B01-98 WT28-B01-100

Table 4.2: Example of WT2G inlinks file.

4.1.1 Topics

The topics (queries) considered for the evaluation of the dataset are 50, formulated with the right criterion and numbered from 401 to 450.
Each of them, so, represents a different informative need, which is submitted to the *Temporal PageRank* system, so that it may try to satisfy it. Each topic, also, is characterized by some useful features, such as:

- **Number**: unique identifier.

- **Title**: significant phrase that represents and indicates the topic of interest, in a short and concise way.

- **Description**: more accurate description of the informative need to be sought within the web pages.

- **Narrative**: text inherent the informative desire, even longer and closer to natural language than the previous.

Below, within a summary table, some of the 50 topics considered for the recovery of the web pages contained inside the collection are shown, pointing out, for each of them, the own characteristics, as well as just announced.

Tag	Content
Numer	401
Title	foreign minorities, Germany
Description	What language and cultural differences impede the integration of foreign minorities in Germany?
Narrative	A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.
Number	402
Title	behavioral genetics
Description	What is happening in the field of behavioral genetics, the study of the relative influence of genetic and environmental factors on an individual's behavior or personality?
Narrative	Documents describing genetic or environmental factors relating to understanding and preventing substance abuse and addictions are relevant. Documents pertaining to attention deficit disorders tied in with genetics are also relevant, as are genetic disorders affecting hearing or muscles. The genome project is relevant when tied in with behavior disorders (i.e., mood disorders, Alzheimer's disease).
Number	403
Title	osteoporosis
Description	Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).

Narrative	A relevant document may include one or more of the dietary intakes in the prevention of osteoporosis. Any discussion of the disturbance of nutrition and mineral metabolism that results in a decrease in bone mass is also relevant.
Number	404
Title	Ireland, peace talks
Description	How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?
Narrative	Any interruptions to the peace process not directly attributable to acts of violence are not relevant.
Number	405
Title	cosmic events
Description	What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?
Narrative	New theories or new interpretations concerning known celestial objects made as a result of new technology are not relevant.
...	...
Number	446
Title	tourists, violence
Description	Where are tourists likely to be subjected to acts of violence causing bodily harm or death?
Narrative	A relevant document must contain accounts of known harm to tourists. Evidence of single, isolated incidents are not relevant.
Number	447
Title	Stirling engine

Description	What new developments and applications are there for the Stirling engine?
Narrative	Any discussion of new developments and applications of the Stirling engine (also known as the Stirling cycle) are relevant.
Number	448
Title	ship losses
Description	Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.
Narrative	Any ship loss due to weather is relevant, either in international or coastal waters.
Number	449
Title	antibiotics ineffectiveness
Description	What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?
Narrative	To be relevant, a document must discuss the reasons or causes for the ineffectiveness of current antibiotics. Relevant documents may also include efforts by pharmaceutical companies and federal government agencies to find new cures, updating current testing phases, new drugs being tested, and the prognosis for the availability of new and effective antibiotics.
Number	450
Title	King Hussein, peace
Description	How significant a figure over the years was the late Jordanian King Hussein in furthering peace in the Middle East?

Narrative	A relevant document must include mention of Israel; King Hussein himself as opposed to other Jordanian officials; discussion of the King's ongoing, previous or upcoming efforts; and efforts pertinent to the peace process, not merely Jordan's relationship with other middle-east countries or the U.S.

Table 4.3: Some topics present inside the WT2G collection.

4.1.2 Query relevance judgments

The query relevance judgments (qrels) made available by the *WT2G* collection, however, are used to verify the quality of the outcomes and the achievements reached by *T-PR*, the Web Information Retrieval system proposed here, in accordance with the first defined and treated topics.
For each of these latter, in fact, its reports are evaluated with respect to the web pages included within the dataset and this is emphasized, furthermore, through the following parameters:

- **Topic**: the topic number.
- **Iteration**: the feedback iteration (almost always zero and not used).
- **WebPage#**: the official web page identifier that corresponds to the "page-no" field in the web pages.

- **Relevance**: a binary code of 0 for not relevant and 1 for relevant.

In agreement with what defined above, some of the query relevant judgments, used for the analysis in question, are shown in the following table.

Topic	**Iteration**	**WebPage#**	**Relevance**
401	0	WT01-B04-284	0
401	0	WT01-B05-1	0
401	0	WT02-B12-219	1
401	0	WT02-B12-220	1
401	0	WT02-B12-221	1
...	...	...	...
450	0	WT24-B40-1	0
450	0	WT25-B01-77	1
450	0	WT27-B10-341	0
450	0	WT27-B13-76	0
450	0	WT27-B24-355	0

Table 4.4: Some query relevance judgments present inside the WT2G collection.

The web pages not included in the qrels file, are not judged by the human revisor, but they are intented as not relevant.

4.1.3 Extraction of temporal expressions

Before the application of the developed case tests, it was necessary to extract the all temporal information present within the web pages and the queries contained inside the *WT2G* collection.
The identification phase of the temporal expressions in question, quite complex and crucial for the application of the concepts that are at the basis of the designed model, was completed through a NLP tool (Natural Language Processing) widely used by the scientific community in this field, called HeidelTime [3] and presented in [22].
Through the same medium, subsequently, the normalization of the same expressions was performed, in order to, thereafter, properly compare the different temporal expressions found and to correctly use them during the analysis of *Temporal PageRank*. In doing so, therefore, the temporal expressions, unearthed through *HeidelTime*, have been stored in a *TimeML* document type, in which each of them is expressed in the *TIMEX3* standard format, according to which, each one is described through the use of three different parameters, respectively named:

- **Offset**: it indicates the start and end position of the expression contained inside the document.

- **Type**: it indicates whether the expression is of date, time, duration, or set type.

- **Value**: it indicates the normalized value of the captured expression.

In the table below, in addition, an example of identification of a temporal expres-

[3] Accessible via `https://github.com/heideltime/heideltime`

sion by applying *HeidelTime* is described.

Phrase	"The 3 th March 2016, i bought a cat!"
File name	SampleTimeML.dtd
TIMEX3 expression	<TIMEX3 tid="0" type="data" value="2016-03-03"> Yesterday </TIMEX3>

Table 4.5: Part of sample TimeML document.

With regard to the *WT2G* collection, however, the processes of identification and normalization of the temporal expressions, present inside it, has been achieved through the execution of a unique and single operation via the command line, by applying it on each of the existing web pages. Below, to summarize, the general procedure which was then undertaken for each of them is shown:

- **Identification and normalization**, performed invoking the Java class called *standalone.jar* and specifying, as parameter, the path of the file of reference (fixed to *webPage.html* in this example), within which to analyze the temporal expressions:

```
java -jar de.unihd.dbs.heideltime.standalone.jar
-t webPage.html
```

As a result of this operation, all temporal data regarding the structure and the composition of the used collection are captured and properly stored.

However, nertheless the temporal expressions are widely present inside the web pages under analysis and in the links that connect them, they are almost absent in the reference topics. To be precise, these features are below point out through two summary table, in which their presence is in detail indicated.

	Temporal aspects
Web pages	[Internal time] 99,8% have at least one temporal expression
Links	[External time] 100%, have at least one temporal expression
Queries	[Internal time] 4%, have at least one temporal expression (0% in title, 2% in description, 2% in narrative field)

Table 4.6: Presence of temporal aspects in WT2G collection.

	Feature
Concentration	the majority of the time is between the years 1996 and 1997, nevertheless more than a century it is covered
Temporal expressions for pages	the web pages present an average of 5 temporal expressions
Temporal expressions for link	the links present an average of 1 temporal expressions
Temporal expressions for queries	the queries present an average of 0,04 temporal expressions

Table 4.7: Distribution of temporal aspects in WT2G collection.

4.2 Test configuration

In agreement with what previously mentioned, each of the factors present in *Temporal PageRank* adopts a concept of "relevance" different from the others, for which, consequently, with respect to it, each factor differently order and evaluate the web pages present within the collection of reference.
In the following sections, so, the contribution of each factor inside *Temporal PageRank* is presented and treated.

4.2.1 Okapi BM25 evaluation

The operation of evaluation of the *TREC* collection through the Okapi BM25 model, which might result complex, is made extremely simple by the Terrier platform, first introduced and presented.
The latter provides, in addiction, the possibility of building a web interface for its search engine, to be used to facilitate the search for data within the considered dataset. Nevertheless, this option was not adopted, preferring to proceed with the analysis using only the command line, deeming it sufficient for the purpose to be achieved.
As suggested by the same Terrier, also, the used operating system was Ubuntu [4] (version 15.10) [5], based on kernel linux, for a better compatibility.
To properly use the platform and to ensure that, through the Okapi BM25 model, the collection of interest is evaluated, some very important and preparatory operations are performed, which involved, among other things, the modification of certain parameters within the *terrier.properties* file.
In relation to the *WT2G* dataset to be evaluated, the following steps were, consequently, followed:

[4] Accessible via `http://www.ubuntu.com/`
[5] Accessible via `http://www.ubuntu-it.org/download`

- **Initial setup**, the first performed operation was carried out to complete the initial setup of the platform, specifying the full path of the *TREC* data collection to be analyzed, which in this case is the *WT2G* and is located in "*/local/collections/WT2G/*":

```
bin/trec_setup.sh /local/collections/WT2G/
```

- **Preparing the collection**, the rebuild of the *collection.spec* file is executed, containing the list of the all files present inside the collection, to allow, subsequently, to be able to index the dataset:

```
find /local/collections/WT2G/ -type f | sort | grep
-v info > etc/collection.spec
```

- **Selecting the model**, among the several available, as earlier anticipated and explained in detail, the selection of the Okapi BM25 model was specified:

```
echo trec.model=org.terrier.matching.models.BM25
>> etc/terrier.properties
```

- **Selecting the topics**, the topics to be used for the analysis are indicated. They are, as already said, 50 and are numbered from 401 to 450:

```
echo trec.topics=/local/collections/WT2G/info/
topics.401-450.gz >> etc/terrier.properties
```

- **Selecting the query relevance assessments**, after indicating the topics, the query relevance assessments are brought to the platform:

```
echo trec.qrels=/local/collections/WT2G/info/
qrels.trec8.small_web.gz >> etc/terrier.properties
```

- **Indexing**, one of the most important steps is that relative to the indexing of the collection, on the base of the previously constructed *collection.spec* file:

```
bin/trec_terrier.sh -i
```

- **Running**, after indexing the collection, the topics are performed on it. To do so, it is necessary to specify a parameter, called c (as suggested it is used a value equal to 0.23), valid to indicate the value for the term frequency normalization. Having done this, the execution can be carried out:

```
bin/trec_terrier.sh -r -c 0.23
```

- **Evaluation**, performed the topics, the next step is relative to the evaluation of the results, with a simple command:

```
bin/trec_terrier.sh -e
```

- **Display**, finally, it is possible to show the obtained results regarding the used metrics:

```
tail -1 var/results/*.eval
```

4.2.2 PageRank evaluation

The entire algorithm used for the PageRank evaluation is described in a single file, written in Python, simply called *PageRank.py*.
Its execution, then, is done via the command line and, as stated by the same author who has developed it, there are no limitations due to the operating system used to fulfill this task.
Being, as already described, a version of PageRank developed specifically for the *WT2G* collection, to evaluate the importance of the web pages contained within it, there was no need to consider addictional measures to run the model in question.
According to what was said before, in fact, it is not necessary to specify particular parameters during its execution. The only modifiable parameters, to be precise, are those related to the damping factor, the number of ranked results to show and the so-called "perplexity".
For the tests carried out, the value of the factor d, or else of the damping factor, has been left the same of the default one, equal to 0.85. The number of useful results to display, however, has been changed from time to time depending on the different

evaluation needs. Also the perplexity was left equal to the default value, or else to 4. This last parameter, it must be said, is very important for the definition of convergence within the algorithm. Being equal to 4, it means that the model reaches the convergence in the case in which, for each node, it is obtained the same PageRank value for 4 consecutive times.
The unique operation, executed via the command line, is so the following:

- **Setup and execution**, after calling the *PageRank.py* file, some addictional parameters were indicated, such as the full path of the *WT2G* collection, the value of the damping factor (fixed to 0,85) and the numer of the most important results to show (in the example below this value is equal to 10):

```
./PageRank.py /local/collections/WT2G/ 0.85 10
```

4.2.3 Temporal PageRank evaluation

As already mentioned, to search for the best configuration of *Temporal PageRank* and to enact, at the same time, its more functional and efficient variant, with respect to the adopted *WT2G* collection, the numerous tests that have been taken into consideration have involved the variation of all the parameters that regulate and manage their composition.
The temporal models were entirely implemented in *Python*, for which, for their use, it was necessary to formulate a precise command through the command line, enhanced by a series of parameters that allow the customization of the execution of the implemented functions.
Below, therefore, the latters are described and treated, in order to fulfill this task,

studying in detail their meaning:

- **Platform setup**, it consists in the initial configuration of the designed platform and it includes the reading of the all web pages belonging to *WT2G* collection, of the topics used to express the informative needs and of the qrels used to judge and analyze each reported result:

```
-collection /local/collections/WT2G/
```

- **Management of the temporal expressions**, after taking awareness of the composition of the reference database, it consists in the understanding of the temporal expressions existing both in web pages and in topics and, therefore, in their proper management. However, it is necessary to consider that, being a program based on the *Unix time* system to describe the time, after the reading of the temporal expressions, some preparatory procedures are conducted on them, to allow the execution of the temporal model:

```
-temporalExpressionsForPages
    /local/collections/WT2G/TE_Pages.json
-temporalExpressionsForTopics
    /local/collections/WT2G/TE_Topics.json
```

- **(Optional) Parameters setting**, it allows to force the use of certain parameter in the model, enabling the setting of the α value, corresponding to the responsabile for the regulation of the linear combination between the Okapi BM25 model

and modified version of the PageRank method. In the absence of such specification, the program tests each possible configuration, looking for the one that is ideal for the case in question. To search for the best configuration, in fact, this is not considered:

```
-a 0.5
```

- **Execution and results**, it consists in the specification of the parameter interested to the processing of the reported results, manually changed from time to time according to the needs of the studies (in this example, the value 5 indicates that only the first 5 retrieved web pages are taken into account in the calculation of the metrics):

```
-top 5
```

4.3 Assessments

As already anticipated and discussed, the goal to achieve and pursue with the following thesis is relative to the improvement of the results that have been reported over time by the best solutions known to the state of the art and by the tools used by the scientific community, overcoming, then, the limits in which those are unbeaten and encounter again.

After some preliminary phases, the reference dataset has been known and the queries available to be submitted to the proposed Web Information Retrieval system have been taken into vision, for which, the main metrics considered to test the developed proposal are here presented. Through these latter, consequently, each reported result was

interpreted and compared, in order to be able to better understand the strengths and the weaknesses that characterize *Temporal PageRank*.
In addition to the known *precision* and *recall*, already presented in the previous chapters, two additional metrics of classification have been adopted and they are the following:

- **F1**: it corresponds to the weighted harmonic mean between *precision* and *recall* and it is also known as the balanced score between them.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

- **MAP**: it is equivalent, for a set of queries Q, to the average precision score obtained for each of them.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

in which, let, for semplicity, p the *precision* and r the *recall*, the first is expressed as a function of the second and is calculated as its average value over r, in the range from 0 to 1:

$$AveP = \int_0^1 p(r)dr$$

Chapter 5

Results

"In an unbroken avalanche of insignificant information, no one knows where to find information that interest."

— Bernard Werber

In the previous chapters, the first attention has been paid to the discussion about the formulation of *Temporal PageRank* and to the ideas that have allowed its birth and its development.
In a second time, then, the configuration used by the tests prepared to challange and check its abilities has been treated and defined, including how every aspect has been managed, the metrics that have been chosen for it and the targets set as the goal to be exceeded.
In this chapter, instead, the light is given to the different results that the proposal was able to reach and gain, comparing these latter to those obtained and reported by the best solutions currently known to the state of the art and to the scientific community, highlighting the ones that are able to be superior or inferior with respect to those latter. As *baseline* and reference point, then, is considered the first presented and accepted Okapi BM25 method but all the comparison are perfomed also with regard

to the $PageRank_{textual}$ model, the solution that consider the original formulation of PageRank and that does not manage, in any way, the time factor.
For all achieved results, to be clear, the used cut-off levels are those equal to 5, 10, 20 and 1000, considering that the latter (1000) is considered the default cut-off level by the same *TREC*.
Entering in detail, then, after the presentation of the results obtained by the baselines and in accordance with the first announced metrics, the scores found by each other elaborate Temporal PageRank variant will be illustrated, accompanying them with summary charts that will have the purpose to compare the proceeding of the obtained MAP measurements varying the α component.
It must be noted that, considering an α value of combination equal to 0, each version of Temporal PageRank is consequently equivalent to the Okapi BM25 method.

OkapiBM25	Precision	Recall	F1	MAP
Top@5	0.50	0.09	0.15	0.08
Top@10	0.44	0.14	0.22	0.11
Top@20	0.35	0.21	0.26	0.14
Top@100	0.17	0.40	0.24	0.21
Top@1000	0.03	0.59	0.05	0.24

Table 5.1: Results obtained through the Okapi BM25 method.

The first analysis has involved $PageRank_{textual}$, the model that does not consider and manage the temporal aspects. This study is important to understand how the time affects the other assessments and it is useful to provide a point of comparison for the other Temporal PageRank variants that are based on modified temporal versions of the PageRank method.

$\boldsymbol{PageRank_{textual}}$	Precision	Recall	F1	MAP
Top@5	0.50	0.09	0.15	0.08
Top@10	0.44	0.14	0.22	0.11
Top@20	0.35	0.21	0.26	0.14
Top@100	0.17	0.40	0.24	0.21
Top@1000	0.03	0.59	0.05	0.24
α combination	0.01			

Table 5.2: Results obtained using through $PageRank_{textual}$.

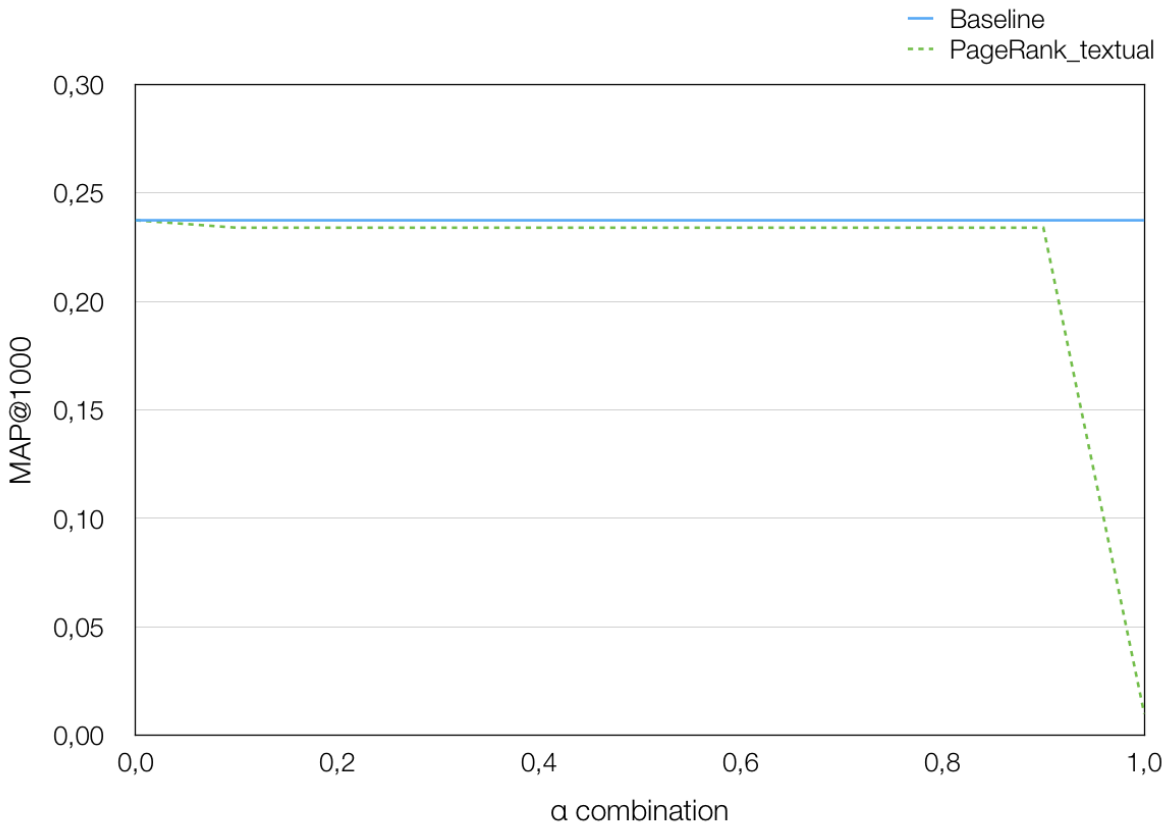


Figure 5.1: MAP measurements of the baseline and $PageRank_{textual}$ for all α combinations.

The use of the PageRank method proves to be unable to bring significant benefits to the Okapi BM25 model, confirming the same results that had been conquered before, with its absence. What was found through $PageRank_{textual}$, moreover, is confirmed by almost all considered α values, as shown in the graph above. Similar experiments concerning the same collection of data, however, had been conducted in the past, as in [12], and had already confirmed that the only PageRank is not able to bring improvements to this web collection, albeit the latter study had considered a different version of the Okapi BM25 model.

The next case taken into account, then, is relative to $TPR_{link-time}$, the TPR variant that consider the time factor only through the meta-information available in the links present in the collection. With respect to what already said, so, $TPR_{link-time}$ compares the link creation time with the current time in which the system is interrogated, rewarding the web pages that possess the most recent links and that appear, therefore, continuously fed in time.

$\boldsymbol{TPR_{link-time}}$	Precision	Recall	F1	MAP
Top@5	0.50	0.09	0.15	0.08
Top@10	0.44	0.14	0.22	0.11
Top@20	0.34	0.20	0.25	0.13
Top@100	0.16	0.39	0.23	0.20
Top@1000	0.02	0.58	0.04	0.23
α combination	0.01			

Table 5.3: Results obtained through $TPR_{link-time}$. Better than baseline: **bold**; worse than baseline: underline.

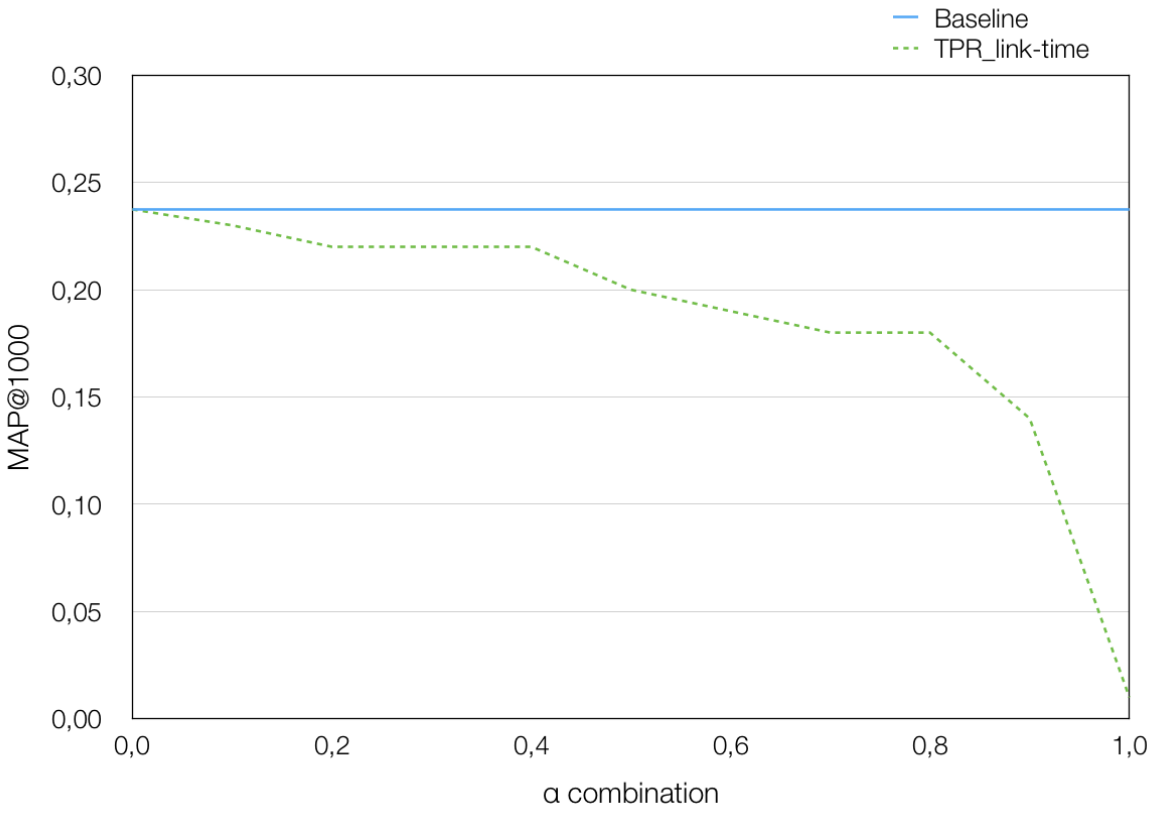


Figure 5.2: MAP measurements of the baseline and $TPR_{link-time}$ for all α combinations.

As can be seen from the table and graph above, in any case the $TPR_{link-time}$ method can prove to be better than the baseline and then $PageRank_{textual}$, although the results differ very little from those obtained by the latter. Considering only the first 5 and 10 retrieved web pages, in fact, with an α combination equal to 0.01, $TPR_{link-time}$ found the same results of the Okapi BM25 method and of $PageRank_{textual}$. At a cut-off level greater than 10, instead, the relevant retrieved web pages decrease and, consequently, the values of precision and recall, and so F1 and MAP, appear slightly lower.
A subsequent experiment, to continue, is inherent to $TPR_{link+query-time}$, the model

that takes into account the temporal aspects extractable from the links, or else recoverable by their creation time, and the internal contents of the queries submitted to the Web Information Retrieval system.

$\boldsymbol{TPR_{link+query-time}}$	Precision	Recall	F1	MAP
Top@5	0.50	0.09	0.15	0.08
Top@10	0.44	0.14	0.22	0.11
Top@20	0.34	0.20	0.25	0.13
Top@100	0.16	0.39	0.23	0.20
Top@1000	0.02	0.58	0.04	0.23
α combination	0.01			

Table 5.4: Results obtained through $TPR_{link+query-time}$. Better than baseline: **bold**; worse than baseline: underline.

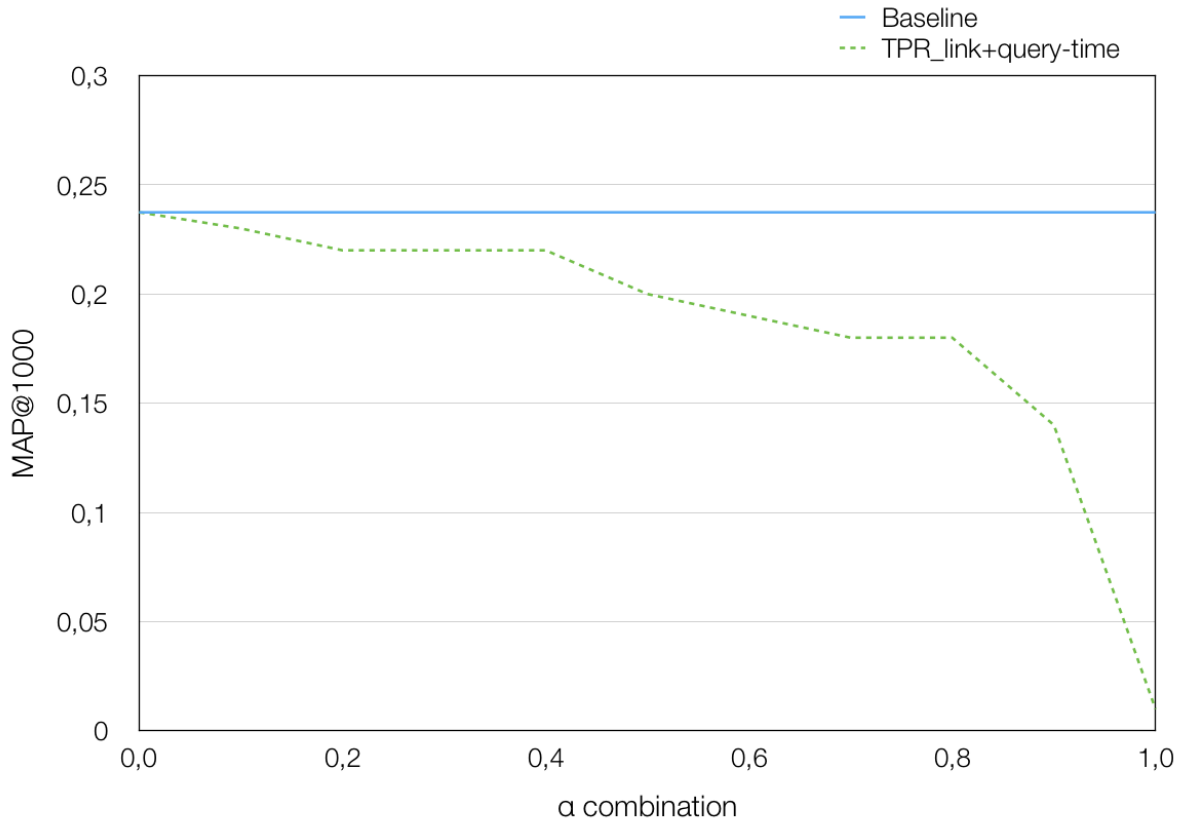


Figure 5.3: MAP measurements of the baseline and $TPR_{link+query-time}$ for all α combinations.

The just conduct analysis shows, as in the previous case, a general trend similar to that reported from the baseline, in particular as regards the first 5 and 10 web pages. The only difference is observable at the cut-off levels greater than 10, from 20 to 1000, for which the new proposal, $TPR_{link+query-time}$, appears to be slightly lower than the Okapi BM25 model and then $PageRank_{textual}$, with regard to the values of precision and recall, and so F1 and MAP.

The next study, to continue, is relative to $TPR_{content+query-time}$, the Temporal PageRank variant that identifies and manages the time factor expressed by the textual content

that characterizes the web pages and the queries used as manifestation of the informative need that we must to satisfy.

$\boldsymbol{TPR_{content+query-time}}$	Precision	Recall	F1	MAP
Top@5	0.50	0.09	0.15	0.08
Top@10	**0.45**	0.14	0.22	**0.12**
Top@20	**0.36**	0.21	0.26	**0.15**
Top@100	0.17	0.40	0.24	0.21
Top@1000	**0.04**	0.59	**0.06**	**0.25**
α combination	0.36			

Table 5.5: Results obtained through $TPR_{content+query-time}$. Better than baseline: **bold**; worse than baseline: underline.

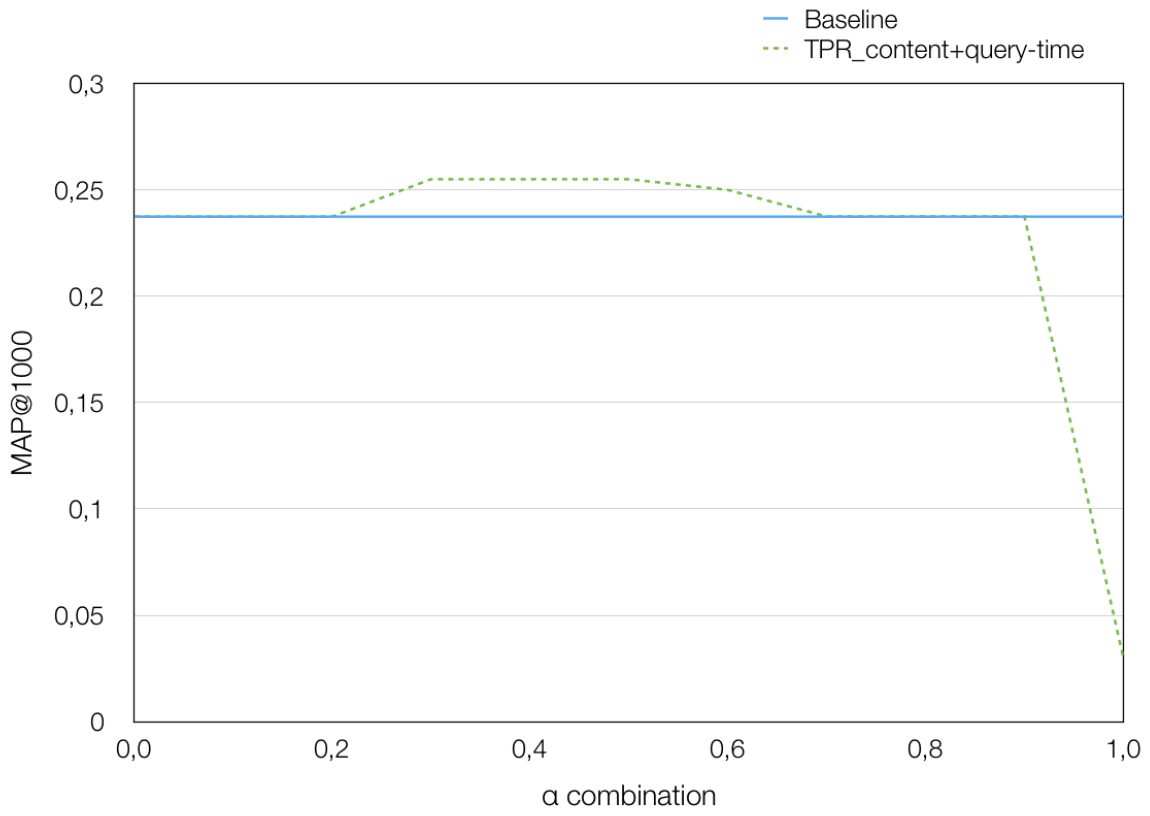


Figure 5.4: MAP measurements of the baseline and $TPR_{content+query-time}$ for all α combinations.

The above reproduced results underline and identify, in contrast to the just treated cases, the first Temporal PageRank variant that is able to rise above the baseline, albeit in a minimal manner. The first advantages to using $TPR_{content+query-time}$ are found in the analysis of the first 10 and 20 retrieved web pages, in fact, the proposal results able to include a slightly higher number of pages considered relevant and it gains slightly higher values of precision and MAP with respect to the baseline and $PageRank_{textual}$. These improvements, in addition, are also found at the cut-off level equal to 1000.
To conclude, then, the last analyzed variant is $TPR_{link+content+query-time}$. Through

it, as already mentioned, the considered temporal aspects are those observable in the links, in the web pages and in the queries used to interrogate the system.

$\boldsymbol{TPR_{link+content+query-time}}$	Precision	Recall	F1	MAP
Top@5	0.50	0.09	0.15	0.08
Top@10	0.44	0.14	0.22	0.11
Top@20	**0.36**	**0.22**	**0.27**	**0.15**
Top@100	**0.18**	0.40	0.24	**0.22**
Top@1000	**0.04**	**0.60**	**0.06**	**0.25**
α combination	0.11			

Table 5.6: Results obtained through $TPR_{link+content+query-time}$. Better than baseline: **bold**; worse than baseline: underline.

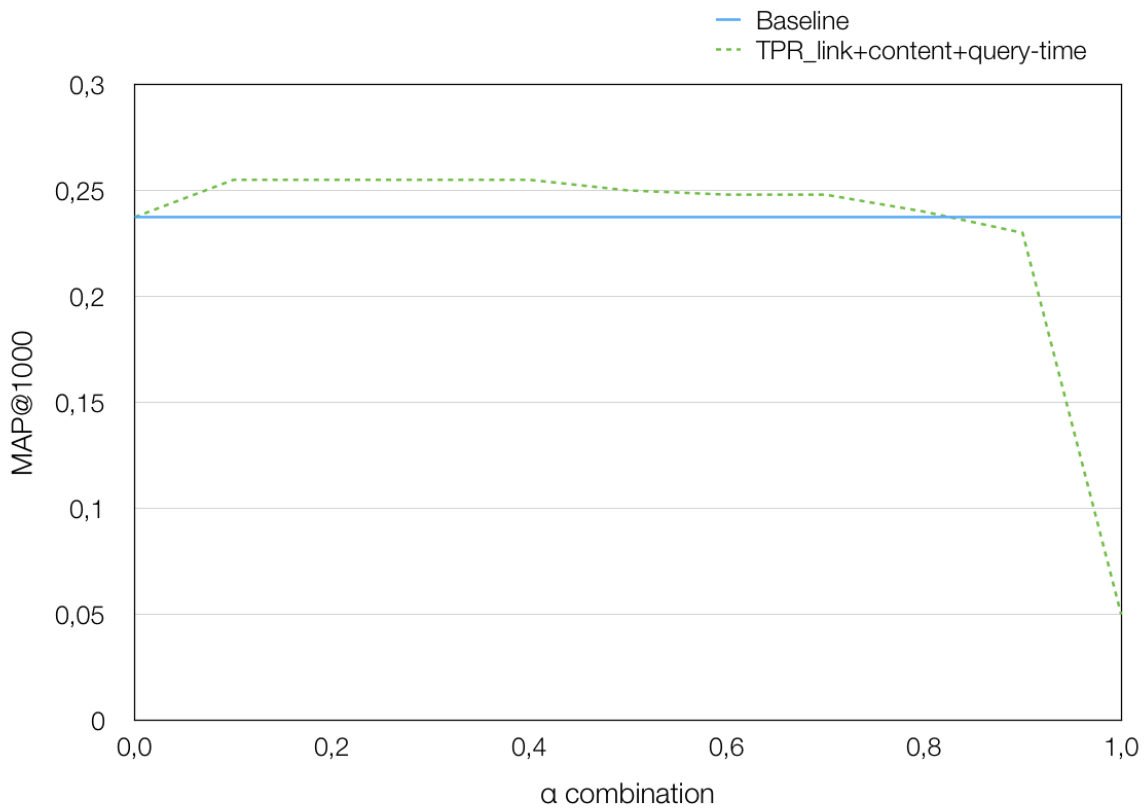


Figure 5.5: MAP measurements of the baseline and $TPR_{link+content+query-time}$ for all α combinations.

As shown, also $TPR_{link+content+query-time}$ manages to bring better results compared to the baseline adopted as reference, albeit in a very reduced manner. For the first 10 retrieved web pages, the model is behaving like the Okapi BM25 method and $PageRank_{textual}$, while with a cut-off level greater than 10, $TPR_{link+content+query-time}$ is able to obtain slightly better values of precision and recall, and so F1 and MAP, thanks to a greater number of relevant retrieved web pages.

To better understand the value of the new proposals presented and analyzed here, two summary charts and a summary table are below provided, with the purpose to compare

the proceeding of the MAP measurement, reported for semplicity at the default cut-off level (1000), and the best configuration of each variant taken into account in this thesis.

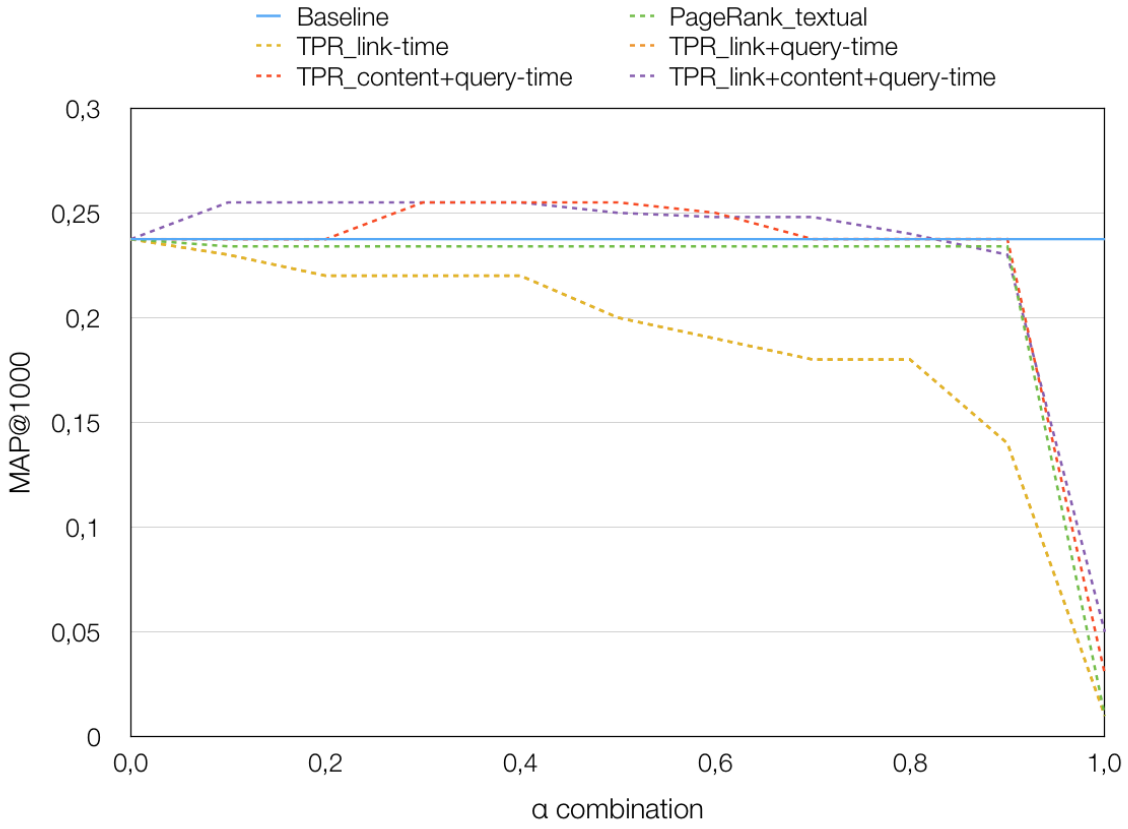


Figure 5.6: MAP measurements of all evaluated methods for all α combinations.

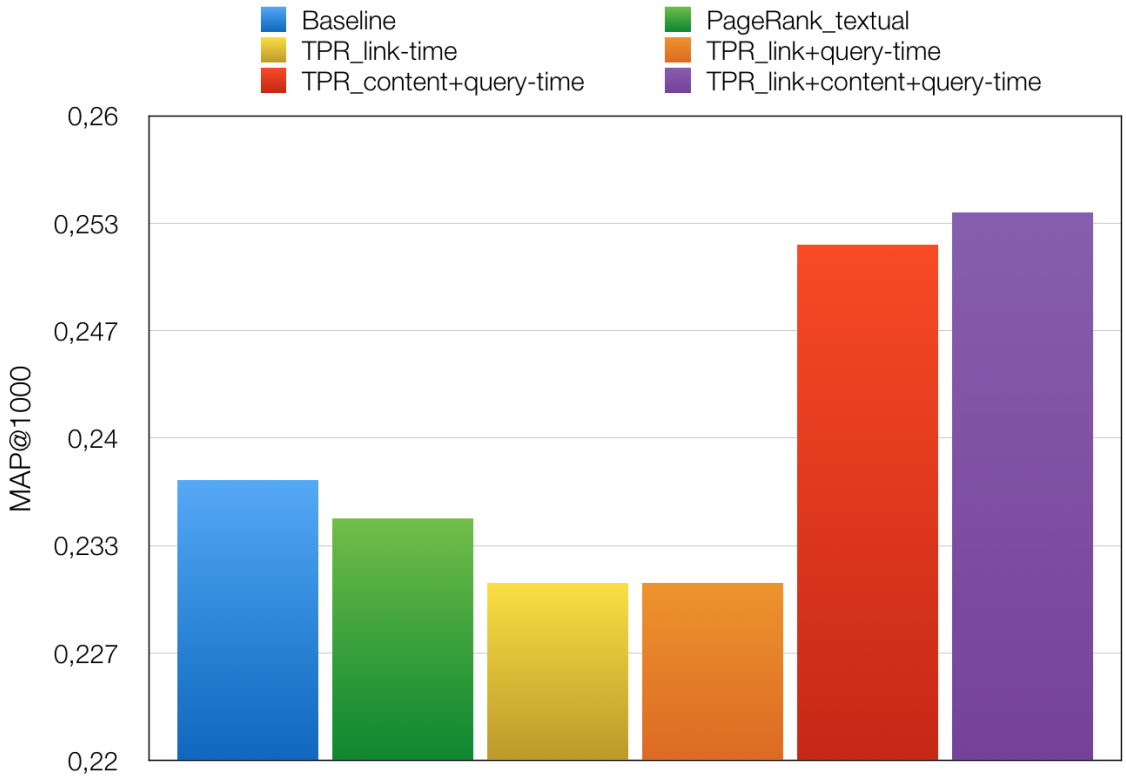


Figure 5.7: MAP measurements for best configuration of Temporal PageRank variants.

Model	P@5	P@10	P@20	R@5	R@10	R@20	MAP
$OkapiBM25$	0.50	0.44	0.35	0.09	0.14	0.21	0.24
$PageRank_{textual}$	0.50	0.44	0.35	0.09	0.14	0.21	0.24
$TPR_{link-time}$	0.50	0.44	0.34	0.09	0.14	0.20	0.23
$TPR_{link+query-time}$	0.50	0.44	0.34	0.09	0.14	0.20	0.23
$TPR_{content+query-time}$	0.50	**0.45**	**0.36**	0.09	0.14	0.21	**0.25**
$TPR_{link+content+query-time}$	0.50	0.44	**0.36**	0.09	0.14	**0.22**	**0.25**

Table 5.7: Results obtained through all considered methods. Better than baseline: **bold**; worse than baseline: underline.

Finally, in light of the performed tests and of the data reported by them, it is possible to deduce that the Temporal PageRank variants that consider the time in the links are not able to be effective and, therefore, to be preferable to the Okapi BM25 model. These results, however, can be attributed to the poor temporal distribution that characterizes the web pages of the used collection. However, the variants that take into account the temporal aspects within the textual content of the web pages are able to achieve better results than the baseline, albeit in a very limited manner. Even in this case, it must be said, the achieved results are mainly due to the temporal concentration of data, which, as already stated, present rather similar time spans between them, making so difficult and complex temporally rank the web pages.

To better verify the importance of Temporal PageRank and the contribution that can be provided in the world of the Web Information Retrieval, therefore, it would be appropriate to test the proposal drawn up in this thesis on other web collections.

Chapter 6

Conclusion and future developments

"There is only one way for the advancement of the science: blame the science already constituted."

— Gaston Bachelard

The development and the progress that characterize our society, as we know, have a major impact on our daily lives. The speed with which events happen and the rapidity with which each of them occurs, however, is directly proportional to the dynamism that involves the dissemination of information relating to them.
Paradoxically, it must be noted, at the growth of the information available to us, corresponds an increase of the difficulty in finding the most suitable and relevant ones between them, that best reflect our needs.
Over time, however, the scientific community has tried to address these problems, solving most of them and allowing all interested users to use tools distributed to facilitate their researches on the web, and not only.

It was born, in fact, the Information Retrieval discipline, which, over time, has attracted more and more researchers, intrigued and fascinated by the potential that could be exploited about the management and the retrieve of information, of different nature, belonging to the various datasets that could be taken into consideration.
In this thesis, therefore, i tried to give a further contribution to the development of the Web Information Retrieval, thereby primarily focusing on the reality of the world of the web, creating a new proposal called **Temporal PageRank**, based on a new concept of relevance by which differently judge the recovery of the pages belonging to the web. In the latter, going more in detail, several factors were analyzed and studied, considering numerous hypotheses, in order to better exploit every aspect that may characterize the researches on the web, starting from the classical textual similarity, already widely adopted by others, passing through the use of the PageRank method, suitably modified to differently understand the importance of a specific web page and concluding with the examination of temporal aspects that allow the identification and management of the chronological and application context that characterize the information.
Then, numerous tests and analysis were prepared to challange *Temporal PageRank* and to better understand its abilities and its weaknesses. The obtained results, however, have shown that the elaborated temporal variants fail to be effectiveness in the all cases taken into consideration, with some exceptions, for which it was recommended the fulfillment of further tests on different web data collections. The results found, to be precise, were explained by the poor temporal distribution that characterizes the information. The majority of the web pages included inside it, as widely expressed, is concentrated in temporal periods similar o very close between them, making difficult, in the adopted collection, use the time component to better rank the web pages. The repetition of these studies on additional collections, therefore, would bring new light on the proposal developed here and would help and facilitate its better understanding

and analysis.

Based on what was said and on the results that have been achieved with the proposal presented here, to continue, some tips and suggestions are below given to allow possible future developments, in order to improve *Temporal PageRank* and help to make it an even better solution.

A first indication, to begin and as already said, might be to conduct new tests on different datasets, so as to understand and discover, through the analysis of the results reported with these latter, the best configuration adoptable by *Temporal PageRank*.

After that, the attention can be moved to the considered variants of the PageRank method and to the temporal models incorporated into the proposal. The main idea might be to improve one of these parts, taking into analysis, if possible, not treated aspects, such as the study of the possible prediction of the PageRank scores that the web pages can assume, in order to pushing up the solution, here present, towards most important and considerable results.

Also, it might investigate the usefulness of exploring spatial information, to favor the recovery of the most spatially relevant information.

Finally, to conclude, a further suggestion might be to consider a new proposal related to the factor responsible for the management of the textual similarity, by ensuring that the latter does not take into account the only occurrence of query keywords inside web pages, but also assess, in addition to this, the possible proximity between terms.

Bibliography

[1] Omar Alonso, Ricardo Baeza-yates, Jannik Strötgen, and Michael Gertz. M.: Temporal information retrieval: Challenges and opportunities. In *In: 1st Temporal Web Analytics Workshop at WWW*, pages 1–8, 2011.

[2] Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel, and Aya Soffer. Trend detection through temporal link analysis. *Journal of the American Society for Information Science and Technology*, 55(14):1270–1281, 2004.

[3] Klaus Berberich, Michalis Vazirgiannis, and Gerhard Weikum. *Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, Rome, Italy, October 16, 2004, Proceeedings*, chapter T-Rank: Time-Aware Authority Ranking, pages 131–142. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[4] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Trans. Internet Technol.*, 5(1):92–128, February 2005.

[5] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 557–566, New York, NY, USA, 2005. ACM.

[6] Brian E Brewington and George Cybenko. How dynamic is the web? *Computer Networks*, 33(1):257–276, 2000.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April 1998.

[8] Matteo Brucato and Danilo Montesi. Metric spaces for temporal information retrieval. In Maarten de Rijke, Tom Kenter, ArjenP. de Vries, ChengXiang Zhai, Franciska de Jong, Kira

Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 385–397. Springer International Publishing, 2014.

[9] Leon Derczynski, Jannik Strötgen, Ricardo Campos, and Omar Alonso. Time and information retrieval: Introduction to the special issue. *Information Processing & Management*, 51(6):786 – 790, 2015.

[10] T. Haveliwala. Efficient computation of pagerank. Technical Report 1999-31, Stanford InfoLab, 1999.

[11] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 517–526, New York, NY, USA, 2002. ACM.

[12] David Hawking, Ellen Voorhees, Nick Craswell, and Peter Bailey. Overview of the trec-8 web track. In *TREC*, 1999.

[13] Ilse CF Ipsen and Teresa M Selee. Pagerank computation, with special attention to dangling nodes. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1281–1296, 2007.

[14] Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. *Internet Math.*, 1(3):335–380, 2003.

[15] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 469–475, New York, NY, USA, 2003. ACM.

[16] Sérgio Nunes, Cristina Ribeiro, and Gabriel David. Use of temporal expressions in web search. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and RyenW. White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 580–584. Springer Berlin Heidelberg, 2008.

[17] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[18] Matthew Richardson, Amit Prakash, and Eric Brill. Beyond pagerank: machine learning for static ranking. In *Proceedings of the 15th international conference on World Wide Web*, pages 707–715. ACM, 2006.

[19] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[20] Stephen E Robertson and Steve Walker. Okapi/keenbow at trec-8. In *TREC*, volume 8, pages 151–162, 1999.

[21] M. Sanderson and W.B. Croft. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451, May 2012.

[22] Jannik Strötgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[23] Jannik Strötgen, Michael Gertz, and Pavel Popov. Extraction and exploration of spatio-temporal information in documents. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 16:1–16:8, New York, NY, USA, 2010. ACM.

[24] Lei Yang, Lei Qi, Yan-Ping Zhao, Bin Gao, and Tie-Yan Liu. Link analysis using time series of web graphs. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007.

[25] P.S. Yu, Xin Li, and Bing Liu. Adding the temporal dimension to search - a case study in publication search. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 543–549, Sept 2005.