Analysis and development of latent semantic indexing techniques for information retrieval

M. Bottello
Saipem Spa, Milan, Italy

Abstract

The majority of information in a company is often useless or not equally interesting to everybody. The set of techniques called Analytical Business Intelligence provides a valid way to tackle the needs of information classification and of knowledge extraction that naturally arises inside a company. This paper proposes a suitable solution to make the digital resources useful, identifiable and accessible, through the creation of a Knowledge Base, i.e. a company-wide Intranet where techniques of Information Retrieval used by the Web are applied in order to reach a higher level of efficiency.
Keywords: wide-Intranet, rank, nodes, arcs, oriented graph, document, Authority, Hub, eigenvalue and eigenvector.

1 Introduction

This article suggests the development of an Intranet as a solution to the problem of information overload in companies. Moreover our analysis points out that an Intranet can be a tool to transform the business information into organizational knowledge. The aim of Intranet is to classify documents without their direct analysis and to exploit the suggestions that the hypertextual nature of the above-mentioned documents provides. The first assumption that had been made is the possibility to adapt to Intranet the techniques developed in the context of Internet, even if the two backgrounds have several affinities. After having examined the differences between the two contexts, we used them to show that the techniques introduced by Kleinberg could be just as valid in different contexts and could lead to considerable results. As far as the methods of document searching are concerned, that through the analysis of the Information Retrieval algorithms that use sophisticated mathematical methods, we were able


WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060201

to show the increase of search efficiency. We have overcome the problems of synonymy and polysemy, and at the same time we have improved the quality of answers to user query, by using the knowledge of the corpus and the possibility of query modification.

2 Intranet Warehouse: the innovative application context

The differences between Web and company wide-Intranet are clear, like for example in the corpus of analysis; the Web pages in Intranet are replaced with texts, the users are many and unknown in the first case, in the second case are skilled and have a well defined profile. Through Intranet we proposed a solution for still opened and not solved Web problems, as for instance the obsolescence of the URLs that point to not existing pages. Then in the Web not all the documents have tag that describes the content: about 20% of the Web pages haven't any title, or when the title exists, it is useless. With an Intranet approach this kind of problems is solved because every document has a title or tag description that is also used to increase the performance of search engine. It is possible to code the content of a document into tags, and this activity is considered as a business best practice by the authors. This development whether guarantees a highest automation in the search and in the data management or it leads to a greater power in the whole knowledge managing.

3 The corpus of the analysis

Given a corpus of documents, the Kleinberg's Information Retrieval techniques allow to reach interesting documents for the users [9]. For each query Q and for each document D, a score (Q, D) is established, that becomes (Q, D, U, C), by adding the users U and the context C. Assume that the author puts in his document links (citations) to other documents, "certifying" implicitly the latter. It is possible to evaluate the authority of a document, by determining the authority of the papers or documents cited. The more authoritative text is the text that has citations of authoritative text. On the other hand the document can increase its authority if it is included in the references of authoritative documents. It is assumed that the documents that cite the same documents treat similar and relevant subjects. The collection, that Information Retrieval techniques analyzed, is made up of documents. The answer to a specific query will be a URL, or directory or folder accessible from all the business units or company areas; the links between documents will be the citations, implicit or explicit, enclosed in every catalogued document. We adopted the following logic: a list of URL is obtained, the documents and the enclosed links or citations are extracted. Strings are associated (context phrase) to every individualized "address", these strings can help to determine the topic of the document. After extracting URL and building the above-mentioned strings, the classification module populates an oriented graph (category tree), where a title is associated to every node. The title consists of only one word or one sentence. The aim of the classification module is to associate weights equal to a number of categories

contained in the Tree for every extracted URL. It possible to assign a URL to a category when the weight of catalog voice is greater than a specific threshold. The weight of each category is calculated by considering the matching between strings of the Tree nodes and the context phrases. The module of pre-classification extracts the names and the noun phrases, from the strings. In this phase the module is called "part of speech tagger", and it recognizes names, verbs, adverbs and adjectives. The comparison between the noun phrases and the titles of categories is made thanks to a lexical database. This method classifies documents after having analyzed their contents. The authors of the text have to follow the proceedings which are strengthened and helped by the structure of the document itself:

- ABSTRACT, a brief summary of the topic;
- CORPUS of the document;
- METADATA and/or KEY WORDS.

The strength of the automatic indexing tool of the citations lies in the ability of individualizing the references to the same paper or text, even if the citations are expressed in different forms and terms. This problem is solved by the authors by standardizing the citations. The document structure facilitates the activity of the search engine, through an identification of the document content that becomes easier. After having recognized the advantages, the authors will follow the main standard scheme; they will contribute to the improvement of Business Technology and to increasing the quality of the query answer. This kind of more structured language solves the problem of the low number of links inside the documents. Also the metadata could give added value, as they are used as simple key words or because they identify the document context. The first useful information for the searching comes directly from the author, the title, the metadata; other information, useful for the matching between query and document, is due to the peculiarity context on which the Information Retrieval techniques are applied. The users are automatically turned to areas or topics tightly to their area of origin. The documents are directly connected to specific business units, operating units and profiles, professional role or position, during the document creation time. The documents are checked and the authors certified other information, like the number of times that the document was read, the creation date of the document useful in the ranking phase, is included. Then the more recent documents are more relevant. The starting-point of the process is the query formulation. We have specific queries, such as specific questions about a topic, (e.g. Which is the up-to-date version of …?). The biggest obstacle is that a small number of documents could include the needed information. The second query–topic is to search specific context information, (e.g. Search information about Java). The third kind is useful similar Internet pages. The user of Intranet, having a good knowledge of its documents in respect to the user of the Web, where the documents are unknown, can formulate his query easily. For this reason we can obtain better results by using specific query. In the case of Intranet the quality of answers is guaranteed directly: the user contributes to refine the query. Therefore the documents are more structured and indirectly the information extraction is facilitated, exploited and certified. Suppose the user

had the necessity to formulate a query that returns documents that treat the same content. In this case, the documents are considered similar because they have the same structure, they analyze the same data or include the same citations, tables or links. A Web risk, in formulating a query–topic, is to return too many pages. In the company wide-Intranet this risk decreases remarkably because the corpus context is known, and users are indirectly linked to the latter and also because the Intranet context is more limited. By the "limit concept" we mean a physical limit: Intranet doesn't have the same size of the Web. Then the Intranet structure is limited because it includes different business areas linked to different users. This kind of structure could increase the quality of the answer.

4 Intranet graph

The company wide-Intranet can be modeled like an oriented graph G = (V, E), where V is the set of nodes and E, set of arcs. In our case, the vertices indicate documents and an arc between pages indicates the presence of a citation of q in p. Suppose there are n pages: P_1, P_2,..., P_n, and P_j points to P_i if in the P_j page there is a link to the P_i page. The graph is defined by a fixed set $V(G)=\{v_1,\ldots, v_n\}$ of elements called nodes and by a set $E(G) = \{e_1,\ldots, e_n\} \subseteq V \times V$ of ordered couples of arcs. Given the arc e = (v, u), the first node (v) is called queue, and arc e is outgoing of v. The (u) is called head, and arc e is ingoing into u. The arc is oriented by node v to node u. This way, we obtain an oriented graph G. A page is more authoritative as there are a lot of links to this page, and its authority depends not only on the total number but also on the authority of the pages that are linked to this one. The Web is an oriented graph, made up of abstract nodes set, and pages, joined by arcs or links, able to codify a large amount of latent. The knowledge of the corpus context reduces the risk of the recording of false data and the risk of links to non existing documents. The main hypothesis of this analysis is that the author of the documents, including a link or citation to another document, gives authority to the second one. Then if a document is cited by another document, the first one contains general information, and the topic is examined in the second one. To define a function that associates a number (rank) to each document, that represents the relevance level of the document after a query, the Latent Semantic Indexing technique was applied [5]. This technique analyzes the textual context and enables to assume that a document is authoritative about a topic if the terms about that topic appear often. It is possible to justify the choice of this model by saying that it works well on multiple queries. By this model it is possible to direct users to the standardization of the document and to the formulation of the query. The graph G(V,E) is represented as a list of adjacent vertices, based on a square matrix n x n, composed of only zero and one, where the generic element (i, j) of the matrix will be equal to 1 if the arc (i, j) of the graph exists, and it is equal to 0 if the (i, j) arc doesn't exist, then:

- $M_{i,j} = 1$ if D_j points to D_i
- $M_{i,j} = 0$ otherwise.

R_i indicates the rank of D_i document and represents the importance of a document. It distributes its importance in a uniform way to the documents that it points to. In order to guarantee the strong connection of the graph, and in order to have two walks oriented inside the graph, one from i-nth to j-nth and the other from j-nth to i-nth, we introduce a "link random" factor in the graph:

$$R_i = (1-\alpha)\, \Sigma_j\, (R_j / N_j) + \alpha\, (1/N)$$

where N is the number of documents. The rank of document i is determined in part, for a fraction $(1-\alpha)$ by the documents that cite i, and for a fraction α the rank is obtained without cost thanks to the presence of arcs from all the documents to document i. Suppose R_i is proportional (through a proportionality constant λ) to the sum of the ranks R_j of the documents D_j that point to D_i.

$$\text{(a) } R_i = \lambda \sum_{j=1}^{n} M_{i,j}\, R_j$$

Considering all the documents of the collection, if R_i are the components of a column vector R n-dimensional, the formula (a) becomes:

$$\text{(b) } R = \lambda\, M\, R.$$

where R is a vector in which every element represents the rank of the relative document. The formulation of the equation (b) means that R is an eigenvector of the adjacency matrix of the Web graph. The rank of a document is calculated by the PageRank algorithm. Not only are the ingoing links used, but also the outgoing ones. U_j represents the number of links in the document D_i that D_j points to. U_j is the sum of the column j-nth of the adjacency matrix. Build a new matrix T dividing every 1 in the column j-nth of M for U_j, and leaving the zeroes unchanged. Then:

$$T_{i,j} = 0 \text{ if } M_{i,j} = 0$$
$$T_{i,j} = 1/U_j \text{ if } M_{i,j} = 1$$

The matrix T is a stochastic matrix, as the sum of every column is equal to 1. This is a transition matrix. We find a number on the position (i, j) of T, which is the probability to move from document D_j to document D_i, after supposing that we are in D_j and we choose at random one of U_j links inside it. It is reasonable to assign a higher rank to the documents that have a higher probability. We can also identify the category of those texts that treat the topic in a general way, the document that more links point to. We define a matrix S, a square binary matrix that has as row and column indices the names of the vertices of the graph. This matrix describes a random process and a rank-vector R is the stationary distribution of the process. The rank-vector R is an eigenvector, it is the normalized eigenvalue of S, that is the sum of its components is 1, with an eigenvalue 1 given by R = S R. R can be obtained by an eigenvector v of S with eigenvalue 1, dividing v for the sum of its components. If we calculate the degrees S of the matrix of transition S^h, the algorithm assumes the fact that the matrix S^h tends, for $h \to +\infty$, towards a matrix where the rows are the same. We have the following theorem [2].

- the limit for h, that tends towards infinitive of the i-nth row of S^h, is the constant row $(R_i, R_i, \ldots, R_i)$ where every element is equal to the i-nth component of R.

- There are a positive constant c and a positive number $\rho < 1$ that, for every i, for every j and for every n, we have, $V = S^h$: $| V_{i,j} - R_i | < c\ \rho^h$.

By this theorem we conclude that if the path followed on the graph is long enough we will arrive to document D_i with probability R_i independently of document D_j from which we've started. The PageRank algorithm [10], here described, is like a stochastic process, or a probabilistic navigator that surfs on the graph and starts from a document i, with probability $(1-\alpha)$ it follows one of links of the current document and with probability α he moves to another document, the rank R_i is the part of time taken by the stochastic navigator in the document i. The PageRank is calculated with great efficiency and the iterative process converges in a few steps. The texts contain explicit and implicit links or citations to authors, texts, Web-sites. We assign a different weight to different categories of links, on the basis of the peculiarity and the precision of the citations. If we consider two kinds of links: explicit link l_e and implicit link l_i. We propose two weights that influence the text ranking. The ratio between them represents the confidence inside the internal documents with reference to the document of random origin (external): w_e and w_i. This subdivision of links is possible because the corpus of analysis is known and because there are different kinds of links between texts. The procedure foresees the counting of the two kinds of links, for every document; in order to build a hierarchy of texts, it is necessary to fix a threshold based on the documents subdivided by authority. A first threshold α refers to a kind of link at a time: if this threshold is overcome the document has a high level of authority. In order not to omit the documents without the first type of links but being authoritative we can define the threshold β. This threshold is to manage the combining of two kinds of links, by using the above mentioned weights. On the basis of the type of user request different weighing of the links could be considered. The weight assigned to the second link category is bigger for specific query, because the aim is to retrieve as many texts as possible that contained the answer of the query. The PageRank considers the probability δ of user carelessness. The algorithm is based on this formula: $PR(A) = (1- \delta) + \delta(PR(T_1) / c(T_1) + \ldots + PR(T_n) / c(T_n))$. PR(A) represents the PageRank of the page A obtained by summing all the PageRank of the pages that point to A, each value divided by the number of links that branch off from that page. The sum value is multiplied by the carelessness factor, a value between 0 and 1 that indicates the probability that a navigator follows really the link. The value is usually equal to 0,85. The value is added to $(1- \delta)$, equal to 0,15. In our case it is improbable that the user doesn't go on in his search passing from a document to other through the links inside the first text. The probability $(1- \delta)$ of following the links in the documents using the matrix of transition T is greater.

5 HITS algorithm (hyperlinked induced topic search)

The aim of the algorithm is to build a list of interesting resources relating to a set of large and well represented topics. There are the following phases: given an item, the system collects a set of documents that contains some terms. This phase

of collection is realized with a search engine that can manage a list of terms like key-words. The search engine presents the result of the query, ordering on the basis of their criteria of scoring. The queries, as a search interface, contain in general a variable number of key-words and the search engine answers by extracting, from its structure, a list of documents that seem to have more relevance (ranking). A fixed number of documents is selected from the results, these documents will form the root set. The root set is extended by adding a document that can be reached starting from the links that are contained in the documents of the root set. In the same way, also the texts that contained links to documents of root set are added. Then the resources, that are one-link-away from root set elements, were retrieved. Then the documents can be divided in two categories: authorities and hubs (study published from Kleinberg in 1998 [9]). The texts are subdivided in two different categories. The "Authority", relevant documents, are "authoritative" information sources for a query, and the "Hub", are texts that "point" to the Authority, list that contain pointers to documents relevant to the item of the query. Every document d receives two ranks. Two scores are assigned to each document, one of hub and one of authority. In the Kleinberg algorithm the weight that is associated to score is unitary. The scores are computed by verifying the number of matching between terms of the item (the same ones used as key-words for the search engine) and the terms of the tag. The aim is to determine the scores of hub and authority for each document that correspond to the eigenvalues of the multiplication of the matrices of weights and of its transposed matrix on the mathematical point of view. The number of iterations, necessary to calculate the scores, is finite. For not negative matrices the value of hub and authority scores converges to two values. By definition, the score of the authority p_a is determined by the hub score associated to the pages that point p. The hub score p_h results by the authority score associated to the pages that are pointed by p. The Hub and the Authority have the "mutual reinforcing" property: a good Authority is pointed by some Hub and a good Hub is pointed by some Authority. The Information Retrieval techniques realize an efficient ranking function, in fact the search by key-word allows one to discriminate the documents that contain key-words from documents that don't contain them. The technique cannot discriminate the high quality document from low quality ones. We observe that the set of relevance criteria, used by search engine, sometimes doesn't correspond with the user needs. An algorithm, to filter the query results by individuating the "authority" and "hub" resources, improves the efficiency of the ranking system. The problem of determining of the "prestigious" documents is solved by supposing that the structure of links contains in latent form the information, necessary to determine a ranking on the basis of prestige. The authority is a document that contains useful information that concerns an item, a hub is a document that contains links to other important texts that concern an item. Let R be the set of document obtained as answer to a query (the index of documents was already build). Be the set S built by adding to the set R, the out-links and the in-links, or the set of documents that points to documents in R and the set of documents that are pointed by the documents in R. Be B(i), the set of documents that points to the document i and F(i) is the set of

documents points by documents i, the algorithm applies in an iterative way two steps I and O. At the first step, we assigned to each page, or document, two different weights, authority and hub that are not negative (initialized with value 1). Then it is computed the score of authority of each text as sum of scores of hub of documents that contain the references. The rule of updating foresees that the weight of authority is the sum of hub weights of documents that point to d, and the weight of hub is the sum of authority weights of documents that d points.

The procedure is: the induced subgraph is obtained starting from $v_{1,\sigma}$, where σ is the query, E a search engine based on the text, t and d are natural number. Suppose R_σ the first t results of E on σ, $S_\sigma = R_\sigma$ for every document p belonging to R_σ, $D^+(p)$, includes all the documents to which p points, and $D^-(p)$ all the documents that point to p, to the set of documents should be added also with high authority the document that more documents point to. If $D^-(p)$ is less or equal to d, all the documents $D^-(p)$ are added or an arbitrary number of d documents from $D^-(p)$ to the subgraph. In this phase, HITS "trusts" the ranking proposed by the search engine. The t documents constitute the root set, that for sure contains the terms indicated in the query, but doesn't always contain an enough number of prestigious pages. For this reason, this second phase foresees an extension of subgraph in order to include all the pages that have a link to the documents of the root set and, in parallel, also the pages reachable by the root set. The following phase, is the document ranking phase, that, in his more simple form, consists of the calculation of the ingoing link of every document, (number of nodes that "link" to it), of the subgraph. This calculation is refined, by observing that near the authorities it is possible also to recognize hub, that have a high outgoing link, a high number of links to other documents. Then, the importance of authority should increase if it is referred by prestigious hub. At the same way the importance of hub should increase if refers good authorities. The HITS algorithm, using the relationship between hub and authority updates the weights that are normalized for each document.

6 The development of Intranet

The HITS procedure, if applied in iterative way, enables to arrive at two constant values of authority and hub. It builds a sort of hierarchy between the hubs, by defining some values, established in advance, to be assigned to hubs. This is possible because the procedure enables to establish if the weight was correctly assigned. Therefore we use the HITS algorithm "to confirm" or "to deny" this hierarchy established in advance by authors, through the assignment to some documents of the probability of being retrieved, that is different from the constant starting value. In this way we build explicitly the hub structure that then will be demonstrated and updated in an implicit and impartial way. The "matrix" interpretation of the HITS method is following, fixed A the matrix of the adjacency of the graph G, the generic aij = 1, if the document i points to the

document j, 0 otherwise. Be h = (hi) and a = (ai) the vector of hub and authority ranks, h = A a and a = AT h, replacing and introducing two constant for the normalization we have that $h = c_1 A A^T a$ and $a = c_2 A^T A a$ must be satisfied. For instance, $a_i = \Sigma_{j \to i} h_i$ and $(A^T h)_i$ = row i of $A^T h$ = column i of Ah = $\Sigma_{j \to i} h_i$.
Be B a matrix n x n with eigenvalues $\lambda_{1, \ldots,} \lambda_n$ such as $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$.

W calculates the eigenvalues ω that correspond to eigenvalues λ_i, of the matrix $B = A A^T$ and $C = A A^T$ respectively. B is the matrix of co-citation where b_{ij} are the number of documents that together point to document i and to document j. C is the matrix where c_{ij} represent the number of documents that together are pointed from document i and from document j. Be v a vector with positive components then we have that the limit, for $k \to \infty$, of B^k v, is a vector parallel to eigenvectors calculated previously. It is always possible to write v as:

$$v = a_1\omega_1 + \ldots + a_n\omega_n .$$

$$B^k v = a_1 B^k \omega_1 + \ldots + a_n B^k \omega_n = a_1\lambda^k{}_1 \omega_1 + \ldots + a_n\lambda^k{}_n \omega_n$$

$$= a_1\lambda^k{}_1 [\omega_1 + a_1/ a_2 (\lambda^k{}_1/\lambda^k{}_2) \omega_2 + \ldots + \ldots]$$

For $k \to \infty$, the factors $(\lambda^k{}_1/\lambda^k{}_2)$ go to zero and the limit of $B^k v = \omega_1$. Then it results that h and a converge to main eigenvalues of $A^T A$ and $A A^T$ respectively.

7 Conclusion

The application, the adaptation in a new context and the development of the Information Retrieval techniques, a company wide-Intranet, allowed providing a solution for the company needs, like the extraction of Information wealth from complex database and the knowledge circle inside the company areas. The solution has the objective of carrying the added value to the Organization and to contribute to the achievement of greater competitiveness. The theme of Human Resources training and development is progressively taking on great importance in the definition of development priorities of Economics and Industry. In companies the quality of techniques and the specialized knowledge of Human Resources is assuming a strategic role, and it is considered as one of main requirements to guarantee the competitiveness and the success. The change of Human Resources knowledge in organizing knowledge, available for all everybody by modern solutions, gives considerable advantages in terms of efficiency and acceleration of professional resources growing. It could be demonstrated that in this context the Data Warehouse, with an efficient search system and reporting, is a basic tool for the company database management. The ranking methods calculate the document relevance for a query on the basis of the presence or absence of words included in the query above mentioned. In this paper it is proposed the Latent Semantic Indexing (LSI) method where the search is according to the content and it is not considered as a theme extraction-generalization, but it could correlate terms as co-occurrences or semantic dominium. It was illustrated how this technique can solve, with statistics, problems of texts ambiguity, like synonymy and polisemy, showing a semantic latent structure that in the documents is hidden for the variability with which the words can be chosen. In short, the main advantage is the achievement of a more efficient problem understanding and problem solving and a careful

understanding of the technical terminology and at the same time the increase in value of the individual experience that encourages the "metabolizing" of the organizational innovation by establishing a cultural exchange between different business areas, branches or foreign countries.

References

[1] Arvind Arasu, PageRank Computation and the Structure of the Web: Experiments and Algorithms, 1999.
[2] Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual {Web} search engine, Computer Networks, 1998.
[3] Andrei Broder et al, Graph structure in the Web, Proc. 9th International World Wide Web Conference, 2000.
[4] A. Broder, A taxonomy of Web search, Journal of ACM, 2002.
[5] Emilio Di Meglio, L'analisi delle corrispondenze per il Text Retrieval con il Latent Semantic Indexing Dipartimento di Matematica e Statistica Università degli Studi di Napoli, 2002.
[6] Susan T. Dumais, Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval, Bellcore, 1992.
[7] Brent Fitzgerald, Implementation of an automated text segmentation system using Hearst's TextTiling algorithm, cs224n final project, 2000.
[8] Taher Haveliwala, Efficient Computation of PageRank 1999 Technical report, Stanford University, 1999.
[9] Jon M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, 1999.
[10] Lawrence Page and Sergey Brin and Rajeev Motwani and Terry Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.