©2010 International Journal of Computer Applications (0975 – 8887)
Volume 1 – No. 12

Performance Assessment using Text Mining

Radha Shakarmani
Asst. Prof, SPIT
Sardar Patel Institute of Technology
Munshi Nagar, Andheri (W)
Mumbai - 400 058

Nikhil Kedar
Student, SPIT
903, Sai Darshan
Versova, Andheri (W)
Mumbai - 400 058

Khandelwal
Student, SPIT
B-401, Mahesh Tower, Sector-2
Charkop, Kavdivli (W)
Mumbai - 400 067

ABSTRACT
Existing search engines have many remarkable capabilities; but what is not among them is deduction capability—the capability to synthesize an answer to a query from bodies of information which reside in various parts of the World Wide Web. Web Intelligence is an area of research which attempts to provide this capability.

Here in this paper we use Text Mining-a feature of Web Intelligence to derive information from the unstructured textual data on the web and device the consensus based strategy to business decisions. This will have two fold advantages, one mitigate the risk early and second would provide a support for our understanding and decision making. This concept is explained with an example of evaluating a player's performance based on minute to minute commentary of the match. Parameters such as his position on field (for example in football – defenders, midfielders, forwards and goal keeper), his past performance, his present fitness and form, and such other parameters are considered. Weightage / value for each parameter is decided and information can be derived for analyzing a player's performance. During analysis we view the comments, we read through fan forums, blogs, newspaper reviews on the play, expert commentator views, etc. This is either used as a correction factor to enhance the credibility of the model.

The whole procedure involves four main stages: Web crawling i.e identifying information resources, information retrieval and extraction, text mining and finally converting unstructured data to structured data.

Keywords

Web Intelligence, NLP, Text Mining, Information Extraction, Information Retrieval, GATE.

1. IDENTIFYING INFORMATION RESOURCES

The first phase is to gather information. But information cannot be gathered from the entire web due to its vastness. To reduce the search time static sources (web site) are chosen to extract information. Authenticity, correctness and up-to-date nature of such sources are very important for information retrieval. Apart from these static sources there might be sources which may provide additional information but are unknown to the user. A web crawler (also known as a web spider) is a program or automated script which browses the World Wide Web to find such sources.

The static site which gives the performance of all the players and their ranking / values are identified and relevant information is retrieved. In addition to the performance, information regarding popularity or fan following of a player can be judged from blog sites.

2. INFORMATION RETRIVAL

Information retrieval system identifies the document, pages in the collection which matches the user query. Information retrieval system allows us to narrow down the set of documents that are relevant to the particular problem.

In our example analysis of minute to minute commentary, news flows and blogs will be done using text mining.

As text mining involves applying very computationally –intensive algorithms to large documents collections, it can be limited to support information retrieved. Information retrieval can speed up analysis considerably reducing the number of documents for analysis. As most of the resources used to gather information are static, it is not necessary that all the information in the form of web pages is required for further analysis.

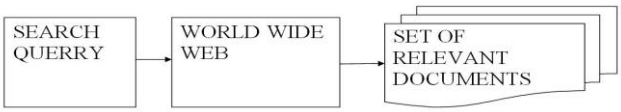


Figure1. Information Retrieval gets sets of relevant documents

In the case study considered, information is derived from the sites where minute by minute description about Football matches is available. The project intends to consider evaluation of players based on expert opinions combined with reviews from common people obtained from blogs. IR systems will allow us to retrieve such documents which will ease the process of text mining. These documents will then be applied to the information extraction systems which are systems used for text mining.

3. INFORMATION EXTRACTION

After short listing the required web pages text mining is applied. The following three phases - information extraction, text mining, converting unstructured data to structured data are closely related to each other.

Information extraction is the process of automatically obtaining structured data from unstructured natural language document. Often this involves defining the general form of information that we are interested in as one or templates which are then used to guide the extraction process. Information extractions rely heavily on the data generated by NLP systems.

The role of Natural Language processing in text mining is to provide linguistic data to next phase. This is done by annotating documents with information like sentences boundaries, parts of speech parsing results. NLP may be deep (parsing every part of every sentence and attempting to account semantically for every part) or shallow (parsing only certain passages or phrases within sentences or producing only limited semantic analysis), and may even use statistical means to disambiguate word senses or multiple parses of the same sentence.

Tasks that Information extraction systems can perform includes:

A) Term analysis: It involves analysis of one or more words, multi word, terms like papers, PDF etc.
B) Named entity recognition: It involves identification of names, dates, expressions of time, quantities, associated percentage, units.
C) Fact extraction: Relationships between entities or events.

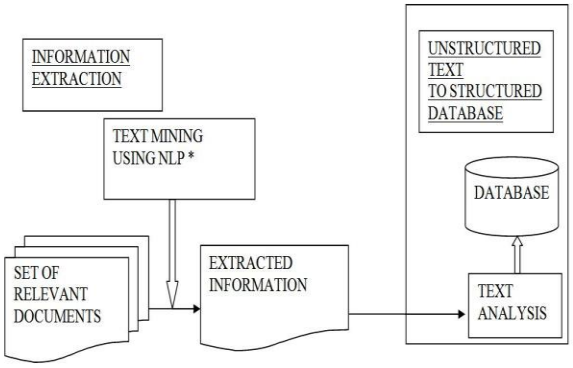


Figure2. Information Extraction and structured to Unstructured data

The data generated during Information extraction phase is structured information derived from unstructured textual data. Information extraction will give information in the form of relationships. Analysis of words, phrases will give us information about an entity and relationship of that entity with other entities. In our case information is used to design a database which can serve as a useful tool to evaluate a player's performance. The database contains attributes such as goals scored, red cards, assists etc for allocating points.

Text mining using NLP is one of the approaches which can be used. Other approaches like Semantic Web and OWL can also be used to provide a solution.

In semantic Web we search for keywords and each keyword is considered a class, if it can be further described (i.e it has attributes) else it is considered an attribute if it has no further description. For example: The classes can be a Player Name and number of goals scored can form the attribute.

In OWL, we create a data dictionary which has all the possible replacements that can be used for a particular attribute .e.g. a player could be referred by his jersey number or by his nick name.

The information that is extracted using NLP also contains similar replacements for the attribute. In this case we can create our own Data Dictionary which can be used as a reference.

The contents of tables are updated at the end of every match. This structured data (e.g.: The number of goal scored by a particular player) is used to determine the performance.

To get information about a player's performance in a particular match we make use of the textual data available as minute to minute commentary on the web. To get information about a player's popularity we can make use of Fan Forum and blogs. Considering performance of a player in a recent match the blogs can give the people's verdict.

Now Natural Language Processing can be used to process this information. Analysis of this information can be used in performance evaluation.

4. IMPLEMENTATION

In our Project the process of Text Mining is implemented using an open source tool kit –GATE (General architecture for Text Engineering).The IE system in Gate is called ANNIE i.e A Nearly New Information Extraction System. (developed by Hamish Cunningham, Valentin Tablan, Diana Maynard, Kalina Bontcheva, Marin Dimitrov and others).

The functioning of ANNIE relies on finite state algorithms and the JAPE (JAVA Annotation Pattern Engineering) language. ANNIE has two parts: Language resources and Processing resources.

Language resources consist of GATE Documents and Corpus. The source of GATE Documents can be specified either by giving the path of a locally stored file on the hard disk or by specifying the URL of a web page. Document formats supported by GATE are XML, HTML, SGML, Plain Text, RTF, Email, PDF and Microsoft Word. A Corpus in GATE is a Java Set whose members are Documents. ANNIE can run only on a corpus.

Processing Resources are responsible for performing text mining on the corpus. There are many plugins available that provide various functionalities. Some of the important ones applicable in our project are Tokeniser, Gazetteer and JAPE Transducer.

The Tokeniser splits the text into simple tokens. There are five types of token – word, number, symbol, punctuation and spacetoken. The aim is to limit the work of the Tokeniser to maximize efficiency, and enable greater flexibility by placing the burden on the grammar rules (JAPE), which are more adaptable.

The Gazetteer Lists used are plain text files, with one entry per line. An index file (lists.def) is used to access these lists; for each list, a major type is specified and, optionally, a minortype. These lists are compiled into finite state machines. Any text tokens that are matched by these machines will be annotated with features specifying the major and minor types. Grammar rules (JAPE) then specify the types to be identified in particular circumstances. Each gazetteer list should reside in the same directory as the index file.

JAPE provides finite state transduction over annotations based on regular expressions. A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The phases run sequentially and constitute a cascade of finite state transducers over annotations. The left-hand-side (LHS) of the rules consist of an annotation pattern that may contain regular expression operators (e.g. *, ?, +). The right-hand-side (RHS) consists of annotation manipulation statements. Annotations matched on the LHS of a rule may be referred to on the RHS by means of labels that are attached to pattern elements. The RHS of the rule contains information about the annotation. Information about the annotation is transferred from the LHS of the rule using the label just described, and annotated with the entity type (which follows it). Finally, attributes and their corresponding values are added to the annotation. Alternatively, the RHS of the rule can contain Java code to create or manipulate annotations. JAPE grammars are written as files with the extension ".jape", which are parsed and compiled at run-time to execute them over the GATE document(s).

In our project we have created Gazetteer Lists consisting of names of players based on the club to which they belong and their position. The Tokeniser is used to annotate the entries in the Gazetteer List. The JAPE rules written use these annotations on the LHS to indentify the player and the JAVA code in the RHS part is to assign points to that player. Different rules are written to indentify various actions like a goal scored, a red/yellow card shown, a save made, a penalty missed, a foul made, a clearance made and so on.

The following two figures describe the implementation

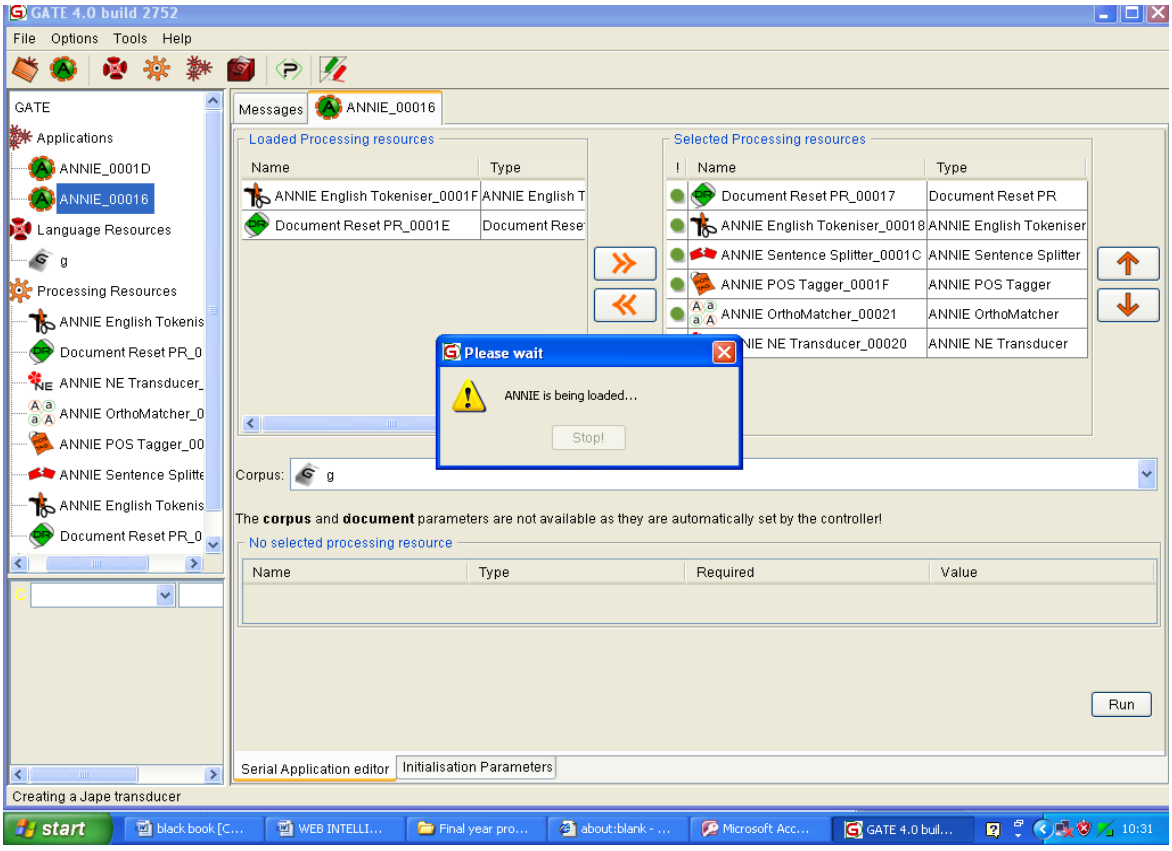
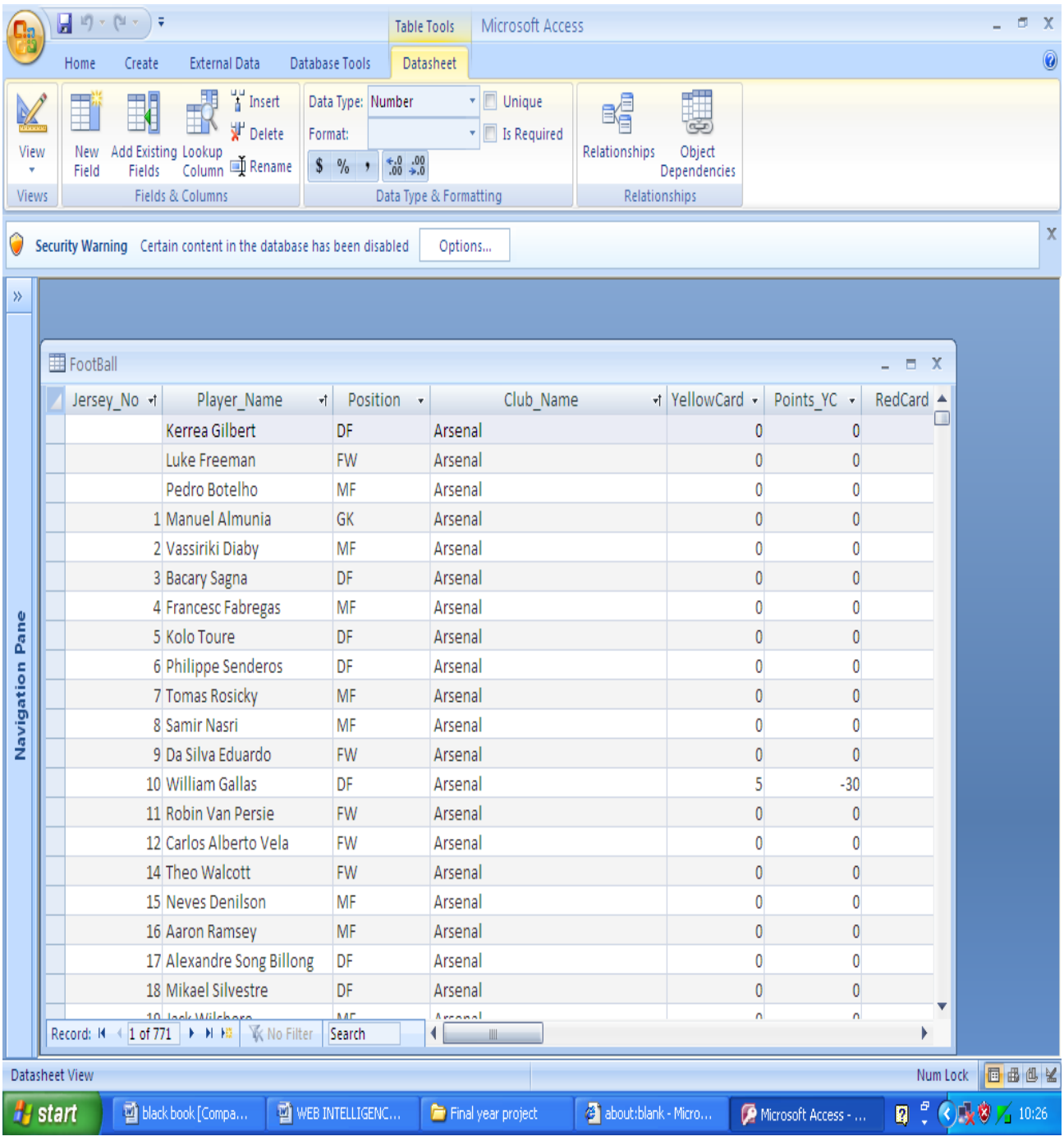


Figure3. A Screen Shot of GATE

Here is a screen shot of GATE which shows the various Language Resources and Processing Resources in GATE and ANNIE being loaded to run on the Corpus



Jersey_No	Player_Name	Position	Club_Name	YellowCard	Points_YC	RedCard
	Kerrea Gilbert	DF	Arsenal	0	0	
	Luke Freeman	FW	Arsenal	0	0	
	Pedro Botelho	MF	Arsenal	0	0	
1	Manuel Almunia	GK	Arsenal	0	0	
2	Vassiriki Diaby	MF	Arsenal	0	0	
3	Bacary Sagna	DF	Arsenal	0	0	
4	Francesc Fabregas	MF	Arsenal	0	0	
5	Kolo Toure	DF	Arsenal	0	0	
6	Philippe Senderos	DF	Arsenal	0	0	
7	Tomas Rosicky	MF	Arsenal	0	0	
8	Samir Nasri	MF	Arsenal	0	0	
9	Da Silva Eduardo	FW	Arsenal	0	0	
10	William Gallas	DF	Arsenal	5	-30	
11	Robin Van Persie	FW	Arsenal	0	0	
12	Carlos Alberto Vela	FW	Arsenal	0	0	
14	Theo Walcott	FW	Arsenal	0	0	
15	Neves Denilson	MF	Arsenal	0	0	
16	Aaron Ramsey	MF	Arsenal	0	0	
17	Alexandre Song Billong	DF	Arsenal	0	0	
18	Mikael Silvestre	DF	Arsenal	0	0	

Figure4.A Screen Shot of the Data Base

Here is a screen shot of the Data Base which shows the allocation of points to a particular based on his performance in the match.

5. OTHER APPLICATIONS

In the above example, information derived was used to provide performance analysis of players in the form of points i.e. structured data. But crawlers along with text mining can support a number of applications. The only things that will change are the information resources and analyzing of information.

To track vote count during presidential elections or to track news about stock market and performance of firms sites for crawling will be different (e.g. news channels websites ,stock exchange web sites) and the analysis instead of database can be dynamic application showing textual updates. One can get information regarding the quarterly reports of a company or about the status of a particular stock and expert opinions on that stock from various websites/news channels all on a single screen. Such an analysis can help a prospective trader to make decision regarding which shares to buy, when to buy and when to sell.

Other applications include applications for analyzing reviews and performance of different products by performing a comparative study, to identify the best product based on the individual needs of a customer. For example if a customer intends to buy a Digital Camera, he would want a comparative report of the various companies to decide the best product for himself. For this websites of companies like SONY, NIKON, etc as well as those of retail outlets need to be crawled.

6. CONCLUSION

The goal of web intelligence is to retrieve information about the customer decision process, customer needs and customer behavior. Retrieving this information gives marketing intelligence the opportunity to improve their predictive models and to create a serious customer view. In this article, we used the example of football league to explain web intelligence using text mining techniques for player assessment.

7. ACKNOWLEDGEMENT

We are very grateful & indebted to our project guide "Prof. Radha Shankarmani" for providing her enduring patience, guidance and invaluable suggestions. She was a constant source of inspiration for us and took utmost interest in our project. We would also like to thank all the IT Staff members for their invaluable co-operation. We are also thankful to all the students for giving us their useful advice and immense co-operation. Lastly we would like to convey our regards to the developers of GATE as well as it's other world-wide users for their constant support and guidance through the online mailing lists.

8. REFERENCES

1] GATE . www.gate.ac.uk. General Architecture for Text Engineering or GATE is a Java software toolkit originally developed at the University of Sheffield since 1995.

2] Web Intelligence: kis.maebashi-it.ac.jp/wi01/ www.web-intelligence.com/

3] Muslea, I. (Ed.). (2004). Papers from the AAAI-2004 Workshop on Adaptive Text Extraction and Mining (ATEM-2004) Workshop, San Jose, CA. AAAI Press.

4] Weiguo Fan, et. al., "Tapping the Power of Text Mining," Communications of the ACM, 49(9), 2006.

5] Tan, A.-H. (1999), "Text Mining: The state of the art and the challenges", in Proceedings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, April, 1999.

6] Intelligence on the Web: www.fas.org/irp/intelwww.html WIN: home WEB INTELLIGENCE NETWORK,smarter.net/

7] J. Srivastava et al., "Web Usage Mining: Discovery andApplications of Usage Patterns from Web Data," SIGKDD Explorations, vol. 1, no. 2, 2000, pp. 12

8] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), "From data mining to knowledge discovery: An overview", in U. Fayyad et al. (eds.) Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, Mass.

9] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), "From data mining to knowledge discovery: An overview", in U. Fayyad et al. (eds.) Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, Mass., 1

10] IEEE 2000b. IEEE Standard for Modelling and Simulation (M&S) High Level Architecture (HLA) –Federate Interface Specification. IEEE Std 1516.1-2000. IEEE Computer Society, New York, NY.