Je-LKS
PEER REVIEWED PAPERS
FOCUS ON SMART, UBIQUITOUS AND MASSIVE LEARNING ENVIRONMENTS
Vol. 11, n.1, 2015
ISSN: 1826-6223 | eISSN: 1971-8829
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

SUPPORTING LEARNING OBJECT REPOSITORY BY AUTOMATIC EXTRACTION OF METADATA

Sergio Miranda
Pierluigi Ritrovato

DIEM - University of Salerno, Italy
semiranda@unisa.it; pritrovato@unisa.it

Keywords: learning; ontologies; semantic representation; Web 2.0

The Learning Objects Repositories are electronic databases able to deliver material on the web allowing instructors sharing and reusing educational units and students accessing and enjoying them. The best way to guarantee these interactions is a good indexing. Each content needs a machine-understandable description able to declare requirements and limits for its right use and to improve any research and delivery action. These descriptions are stored in the metadata. Filling in the metadata is a boring and time-consuming activity but it is very important since it could influence the choice of the best material to deliver. This paper describes a possible methodological approach to automate this activity by extracting metadata directly from the files setting up the learning object itself. In the literature, there are many methods able to automatically characterize the technological aspects of the content, but very few of them are able to provide information about its pedagogical features. The proposed approach tries to draw together

for citations:
Miranda S., Ritrovato P. (2015), *Supporting Learning Object Repository by automatic extraction of metadata,* Journal of e-Learning and Knowledge Society, v.11, n.1, 43-54. ISSN: 1826-6223, e-ISSN:1971-8829

information theory, learning models, statistical analysis and ad hoc heuristics to extract a wide set of fields of the metadata. The results of a first experimentation are particularly encouraging to think about this approach as a solution to support learning object repositories and other platforms having needs to manage wide content storage and huge amount of users with various personal features, devices for interaction and goals as in the MOOCs.

1 Introduction

A learning object repository (LOR) is a kind of digital library where educators may share, manage and use educational resources and learners may find, access and enjoy them. A use of content/assets/resources in a traditional or distance learning environments requires that they have to be found and, for a good searching process, they have to be indexed and well described by means of complete and standard metadata (Miranda & Ritrovato, 2014).

Thus, the faced problem is to automatically extract metadata from the content in order to efficiently and effectively catalogue and reuse it. More in detail, we analysed the innovative platform Intelligent Web Teacher (IWT) (Capuano *et al.*, 2009), that is able to create and deliver courses personalized by respecting the features of the users and to guarantee flexibility in terms of content and applicable learning models. IWT, by leveraging these key points, is more effective than other conventional e-learning solutions because its paths are sequence of learning objects selected as the best way for each different learner to gain her learning goal. The personalization of courses is based on modelling the knowledge domain by means of ontologies and metadata of the content and user profiles in terms of cognitive state and learning preferences constantly updated by means of tests and interaction.

The experimentation described in (Capuano *et al.*, 2008) revealed that the approach of IWT, compared with other solutions, methodologies and models available in the literature, is still valid in both educational and training contexts. The results of the courses personalized by IWT are better (in terms of pedagogical effectiveness (Austin *et al.*, 2010)) than results of the static sequential courses of other conventional e-learning solutions. This approach is automatic but requires a start-up effort to define metadata of learning resources in order to index and appropriately use them. IWT became the core of a new platform that is trying to face one of the main problems of Massive Open Online Courses (MOOCs): the dropout. The main cause of this is the difficulty to guarantee the presence of a one-to-one tutor for many learners. The proposed training platform, in particular, exploits the adaptation and personalization features of IWT to mitigate this cited problem (Miranda *et al.*, 2013). MOOC-solutions and other systems where Open Educational Resources are employed for smart education issues (Marcus-Quinn & Diggins, 2013) have big databases of le-

arning objects and treat huge number of users. These situations mean more content to index and different user-requirements to satisfy, thus the proposed approach may represent a solution that is applicable in a wide variety of cases.

Table 1
METADATA FIELDS, ADMITTED VALUES AND CARDINALITY

	Metadata field	Value	Cardinality
1	Language	{ IT, EN, ES, FR, ...}	1-N
2	Domain/Concept	Text	0-1/1 < = > 1-N
3	File type	MIME Type	0-N
4	Dimension	Number of bytes	0-1
5	Learning resource type	{ SELFASSESSMENT, DIAGRAM, EXAM, EXERCISE, EXPERIMENT, FIGURE, GRAPHICS, INDEX, LECTURE, PROBLEM, QUESTIONNAIRE, SIMULATION, SLIDE, TABLE, TEXT }	0-N
6	Duration	HH MM SS	1
7	Interactivity type	{ Expositive, Active, Mixed }	0-1
8	Interactivity level	{ Very low, Low, Medium, High, Very high }	0-1
9	Difficulty	{ Very easy, Easy, Average, Difficult, Very difficult }	0-1
10	Semantic density	{ Very low, Low, Medium, High, Very high }	0-1
11	Time to learn	HH MM SS	0-1

2 Metadata

The key points for the content indexing are the metadata. Dublin Core, IMS Metadata (defined by IMS Global Learning Consortium) and IEEE LOM (Learning Object Metadata defined by Learning Technology Standardization Committee, LTSC dell' IEEE) are the ones most commonly used. Dublin Core is a general purpose standard, the other ones are learning-specific and, unlike the first one, have a hierarchical structure. Each element may be mandatory, optional or conditional. Some attribute may have, as its value, a list of terms coming from a specific dictionary instead of a single value. IWT uses IMS Learning Object Metadata. In our work, we paid attention to the metadata fields showed in the following Table 1 that represents the minimal set of features used for the personalization.

3 Literature review

Many approaches and models were aimed at the automatic extraction of metadata fields directly from the content. The recent works (Atkinson *et al.*, 2013) took contexts into account, the other approaches (Dharinya & Jayanthi,

2013) were aimed to identify objectives and prerequisites. In our work, objectives are clear, but we need details on aspects to describe the content itself as treated by other techniques. Among them, the approach developed by National Library from New Zealand, by HATII, University of Glasgow[1], and other models[2] seem to be the most promising. All of them are able to treat a wide set of content type like pdf, image, Microsoft Office document, audio file, etc. but they produce technical information on the file itself and on the application to deliver and benefit from (Greenberg, 2004). However, they are unable to extract useful data for learning issues as: *Learning resource type, Interactivity type, Interactivity level, Difficulty, Semantic density, Time to learn*. For that, there some other models developed by Ronsivalle (Ronsivalle *et al.*, 2009), Bloom, Anderson, Marzano and Romiszowski[3] and others (Sagi *et al.*, 2009). These models, while very effective in describing educational features of the content, have the drawback that they work a-posteriori. This means that these models are able to characterize the content itself after its delivery by leveraging statistical surveys on time and other parameters, by inferring complexity and semantic density and by gathering the amount of the transferred information. However, their validity is a reference point to develop our heuristics.

4 Automatic metadata extraction

By analysing technical details of the content and applying principles of the Information Theory (Shannon, 1993), the proposed approach aims at producing metadata. This process needs that each Learning Object (LO) has a semantic description that specifies which concept of which domain the content explains (i.e. the field 2 *Domain/Concept* of the Table 1). After that, let us consider the set of files included in the LO. Each of them has a file extension and a known number of bytes. This allows evaluating the field 3, *File type* and the field 4, *Dimension*. To find the right MIME type to fill in the field 3, we refer to a schema[4].

[1] Humanities Advanced Technology and Information Institute (HATII), University of Glasgow http://www.digitalpreservationeurope.eu/publications/briefs/semantic%20metatada.pdf

[2] Metadata Extraction, http://wiki.alfresco.com/wiki/Metadata_Extraction; Milena Dobreva, Yunhyong Kim and Seamus Ross. Automated Metadata Extraction http://www.dcc.ac.uk/resources/curation-reference-manual/chapters-production/automated-metadata-extraction; Apache Tika. http://tika.apache.org/

[3] Fun with learning taxonomies of Bloom, Anderson, Marzano and Romiszowski, http://gramconsulting.com/2009/02/fun-with-learning-taxonomies/

[4] MIME Types: http://www.asciitable.it/mimetypes.asp

Table 2
HEURISTIC RULES TO FILL IN THE FIELDS 5, LEARNING RESOURCE TYPE, 7, INTERACTIVITY TYPE AND 8, INTERACTIVITY LEVEL FROM THE MIME TYPE

	INPUT	OUTPUT		
Name	MIME Type	Learning resource type	Interactivity type	Interactity level
Rule 1	application/excel	TABLE	MIXED	MEDIUM
Rule 2	application/msword	TEXT	EXPOSITIVE	VERY LOW
Rule 3	application/ mspowerpoint	SLIDE	EXPOSITIVE	LOW
Rule 4	application/pdf	TEXT	EXPOSITIVE	VERY LOW
Rule 5	application/ wordperfect*	TEXT	EXPOSITIVE	VERY LOW
Rule 6	"application/*" (except app.1-5)	SIMULATION	ACTIVE	VERY HIGH
Rule 7	text/*	TEXT	EXPOSITIVE	VERY LOW
Rule 8	audio/*	LECTURE	EXPOSITIVE	VERY LOW
Rule 9	video/*	LECTURE	EXPOSITIVE	LOW
Rule 10	image/*	FIGURE	EXPOSITIVE	VERY LOW
Rule 11	*world*/*	SIMULATION	ACTIVE	VERY HIGH
Rule 12	*message/*	TEXT	EXPOSITIVE	VERY LOW
Rule 13	*conference/*	LECTURE	MIXED	MEDIUM
Rule 14	drawing/*	GRAPHICS	EXPOSITIVE	VERY LOW
Rule 15	chemical/*	DIAGRAM	EXPOSITIVE	VERY LOW
Rule 16	model/*	SIMULATION	ACTIVE	VERY HIGH
Rule 17	paleovu/*	GRAPHICS	EXPOSITIVE	VERY LOW
Rule 18	*/*metafile*	INDEX	EXPOSITIVE	LOW

In case of more than one file, we load all related MIME types in the field because it has 0-N cardinality. By using the loaded MIME types, we may estimate the fields 5, *Learning resource type*, 7, *Interactivity type* and 8, *Interactivity level*. We analysed the files and the features of related LO in order to create the heuristics to apply. The Table 2 shows the identified set of rules. The field 5 has *0-N* cardinality, thus it contains a different type for each different file. The field 7, instead, has *0-1* cardinality, thus, in case we have more than one file, we defined a heuristic to apply iteratively to all files of the object pair by pair. The field 8, similarly with the previous one, has *0-1* cardinality, thus, in case we have more than one file, we defined a heuristic to apply iteratively to all files of the object pair by pair. To calculate the *Duration* of each LO, instead of access to its content, we tried to speed up this operation by estimating it directly from the type of the files. For text files, we adopted an approach similar

to "Writer Services"[5] and were able to estimate the number of words in the content from the dimension of its file. We considered many formats (txt, html, pdf, doc, rtf, docx …) and tried it on documents whose number of words was a-priori known. Fig.1 shows how the number of Bytes for the different treated formats is related to the number of words in a document. Thus, the number of words grows when the size grows.

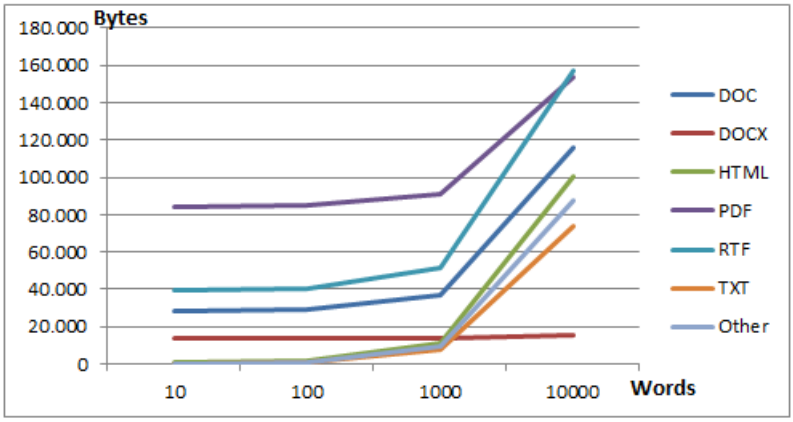

Fig.1 - Growth of the dimension for each file format depending on the number of words

By using these trends, we elaborated an estimation function able to estimate the number of words in a content from the dimension and the format of its file. We conducted similar analysis on different file formats in order to estimate as *Brandon Hall* (Clarey, s.d.), the time to spend on figures, graphics and diagrams. This analysis gave us a first result on the correlation between the time and the image quality. It implies, by overlooking any compression on the format, that the time may be related to the number of Bytes. For the objects having a high-level interaction like simulations, experiments and exercises, we made empirical evaluation on some statistical bases in order to estimate the time parameter starting from the dimension of this kind of LO. Moreover, for the lectures (audio, video resources and synchronous live events) the time parameter has the same value of the duration of the object. We have to underline that the *Duration* is not equal to the *Time to learn*. To estimate it, we need to describe more in detail the LO by evaluating its both metadata fields *Semantic density* and *Difficulty*. The time a user may spend in reading a text grows as well as the higher *Semantic density* or the higher *Difficulty* is. In fact, a higher *Difficulty* implies a more time required for the comprehension. Similarly, a

[5] Word count to page, http://www.writersservices.com/wps/p_word_count.htm

higher *Semantic density* implies more time required for the comprehension. In our approach, each LO should be referred to a set of concepts. This is what usually happens in IWT to index learning material: after a semantic description of a specific domain by using lightweight ontologies (Capuano *et al.*, 2009), each LO is linked to concepts (just one or a set of them) it explains. By applying the principles of the Information Theory (Shannon, 1993), we can adopt the set of concepts as the amount of information in the LO. Thus, for a given set of concepts, a higher dimension means a lower *Semantic density* and vice versa. This allows us calculating the *Semantic density* of a LO ds_c as the ratio between the number of concepts n_c and the dimension of the LO itself (the sum of the dimensions d_f of all files *f*):

$$ds_c = \frac{n_c}{\sum_f d_f} \tag{1}$$

This is the amount of information used to treat the set of concepts *c*. By taking into account the models (Ronsivalle *et al.*, 2009) we may evaluate the *Difficulty* by mean of a relationship between the concepts treated in the LO and its *Duration*. For a given set of concepts, the less *Duration* corresponds to the higher *Difficulty* and vice versa. Thus, we may calculate the *Difficulty* d_c as the ratio between the number of concepts n_c and the total *Duration* (sum of the duration t_f of all files *f*):

$$d_c = \frac{n_c}{\sum_f t_f} \tag{2}$$

The fields 9, *Difficulty* and 10, *Semantic density*, have a *0-1* cardinality and a limited set of possible values (as showed in the Table 1). Of course, this opens another debate: which is the right dimension for a concept? Which is the right time for a concept? In other words, what does medium *Semantic density* mean? What does medium *Difficulty* mean? To find the best answers to these questions and point out the range of value to use, we performed a statistical analysis on available LOs. After this, the last field to estimate is the 11th: the *Time to learn*. This depends on both *Difficulty* and *Semantic density*. A higher *Difficulty* of a LO on a concept means more time to understand the concept. Similarly, a higher *Semantic density* of a LO means more time to spend for learning. Thus, we may calculate the time to spend on a LO (tlo_c) by multiplying its *Duration* (d_c), with the factor related to the *Difficulty* (fd_c) and the factor related to the *Semantic density* (fds_c):

$$tlo_c = d_c \cdot fd_c \cdot fds_c \tag{3}$$

5 Experimental results

In order to verify the quality of the proposed approach, we analysed a database containing more than 2600 LOs on subjects related to *Mathematics* and *Computer science* and another database containing more than 800 LOs on subjects related to some *procedures of the large-scale retail trade*. By means of the IWT platform, we included these LOs in both simple and personalized learning courses and involved more than 2500 users in different periods. We tracked all times spent by users and we collected their perception on *Difficulty* and *Semantic density* by asking them to fill in an on-line survey created on the base of the principles of Ronsivalle on Instructional Design (Ronsivalle *et al.*, 2009). For a fixed topic (e.g. "*Fundamentals of computer science, the concept of IF*"), we asked users to give their feeling on how they perceived *Difficulty* of the content and *Semantic density* of what they saw. They submitted their answers on a web form in terms of Likert responses.

For the *Difficulty*, we applied these interpretations: *Very easy* does mean knowledge on treated concepts (remember the concepts that have been presented); *Easy* does mean comprehension on treated concepts (clear concepts and feeling of knowing them); *Average* does mean application of treated concepts (feeling able to apply concepts to practical problems); *Difficult* does mean analysis on treated concepts (feeling able to analyse information critically); *Very difficult* does mean synthesis on treated concepts (feeling able to do abstraction from information and synthesis).

For the *Semantic density*, we applied these interpretations: *Very low* does mean that the amount of information is not sufficient to justify this learning unit; *Low* does mean that the amount of information is the minimum to justify this learning unit; *Medium* does mean that the amount of information is the optimum to justify this learning unit; *High* does mean that the amount of information is the maximum to justify this learning unit; *Very high* does mean that the amount of information is too much to justify this learning unit.

By collecting and interpreting the user's feedback, we refined the estimation models of our approach to fill in the metadata fields automatically. First, we collected all statistical details about *Duration* and *Dimension* of the LO. The considered data set includes 1021 learning objects (about 30% of 3400 available LOs: 24% from Math domain and 37% from Retail domain). The number of involved users is 979 (about 40% of 2500 registered users: 18% from Math domain and 61% from Retail domain). Each learning object has been delivered

to more than one user. The system tracked all the times for all of them. The maximum value is about 7 minutes. The minimum value is about 2 seconds. Among all the values on the considered data set, the average *Duration* is about 3 minutes, the variance is 2.57 and the standard deviation is 1.58. No substantial differences have been observed between the two considered domains. On the same data set of learning objects, we collected all the values of the Dimension parameter. The maximum value is about 2200 KBytes. The minimum value is about 12 KBytes. Among all the values on the considered data set, the average Dimension of the related files that is about 150 KBytes, the variance is 312648.5 and the standard deviation is 550.9. No substantial differences have been observed between the two considered domains.

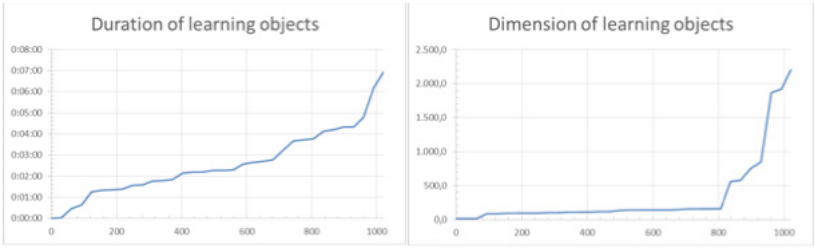

Fig.2 - The distribution of the Duration and Dimension of the learning objects

We considered these values as the most suitable ranges for these both parameters. We correlated this evaluation with the estimation of *Difficulty* and *Semantic density* and we compared this correlation with the feeling of the users.

The following rules derived from the results of this analysis on the *Difficulty* related to the *Duration*: If the *Duration* is more than 5 minutes, then the *Difficulty* is *Very easy*; If the *Duration* is more than 3.5 minutes and less than 5 minutes, then the *Difficulty* is *Easy*; If the *Duration* is more than 2.5 minutes and less than 3.5 minutes, then the *Difficulty* is *Average*; If the *Duration* is more than 1 minute and less than 2.5 minutes, then the *Difficulty* is *Difficult*; If the *Duration* is less than 1 minute, then the *Difficulty* is *Very difficult*.

The following rules derived from the results of this analysis on the *Semantic density* related to the *Dimension*: If the *Dimension* is more than 1500 KBytes, then the *Semantic density* is *Very low*; If the *Dimension* is more than 200 KBytes and less than 1500 KBytes, then the *Semantic density* is *Low*; If the *Dimension* is more than 100 KBytes and less than 200 KBytes, then the *Semantic density* is *Medium*; If the *Dimension* is more than 50 KBytes and less than 100 KBytes, then the *Semantic density* is *High*; If the *Dimension* is less than 50 KBytes, then the *Semantic density* is *Very high*.

Secondly, starting from the statistical collection of data we calculated both the factors related to the *Difficulty* and to the *Semantic density* used to estimate the time to spend on a LO. The most likely influence factors to use are shown in the following figure 3.

		Semantic density	*Very low*	*Low*	*Medium*	*High*	*Very high*
		Semantic density factor for time	*1,2*	*1,3*	*1,5*	*1,8*	*2*
Difficulty	Difficulty factor for time						
Very easy	*1,2*		1,44	1,56	1,80	2,16	2,40
Easy	*1,3*		1,56	1,69	1,95	2,34	2,60
Average	*1,5*		1,80	1,95	2,25	2,70	3,00
Difficult	*1,8*		2,16	2,34	2,70	3,24	3,60
Very difficult	*2*		2,40	2,60	3,00	3,60	4,00

Fig. 3 - The influence factor of the Difficulty and the Semantic Density on the time to learn

We received feedbacks from about 40% of the users on about 30% of the available material; this is not what we aimed at the beginning, but it is enough to refine our estimation approach.

Conclusions

In the contexts of content management, metadata become essential to support sharing and reusing processes. In particular, for the big learning objects repositories, the indexing plays a key role in guarantying good interactions. The success of a learning object repository depends on the specific characteristics that meet the needs of instructors, designers and learners. A learning object should be easily accessed at the exact moment that the design needs it or learning activity calls for it. This may happen by filling in the metadata fields. Usually, this activity is boring and time-consuming. For this reason, we proposed a simple, efficient and quite original approach for the automatic classification of the content of a digital library. We defined a model to extract metadata from the objects by avoiding expensive analysis and applying elaborations directly on the files. This approach has been experimented into an e-learning environment, but it may be used in other contexts where it could be an extension of other existing applications. It could be a plug-in for systems where there are the needs of indexing content to improve search engine performances and allow best matching between available material and customer

requirements. MOOC environments and open educational resources are only some examples of contexts where enterprises may need to index and classify huge amount of documents in order to give best answers to real user needs.

REFERENCES

Atkinson, J., Gonzalez, A., Munoz, M. & Astudillo, H. (2013), *Web Metadata Extraction and Semantic Indexing for Learning Objects Extraction*. Recent Trends in Applied Artificial Intelligence. Springer Berlin Heidelberg, pp. 131-140.

Austin, R., Smyth, J., Rickard, A. & Quirk-Bolt, N. (2010), *Collaborative digital learning in schools: Teacher perceptions of purpose and effectiveness*. Technology, Pedagogy and Education, 19(3), pp. 327-343.

Capuano, N. *et al.*, (2008), *LIA: an Intelligent Advisor for e-Learning*. Emerging Technologies and Information Systems for the Knowledge Society - Proceedings of the World Summit on the Knowledge Society (WSKS 2008), 24-26 September, Volume vol. 5288, pp. 187-196.

Capuano, N., Miranda, S. & Orciuoli, F. (2009), *IWT: A Semantic Web-based Educational System*. Proceedings of the IV Workshop of the AI*IA Working Group on Artificial Intelligence & e-Learning held in conjunction with the XI International Conference of the Italian Association for Artificial Intelligence (AI*IA 2009), Volume AI*IA, 2009, pp. 11-16.

Clarey, J., s.d. *How do you estimate e-learning seat time?*. [Online] Available at: http://www.brandon-hall.com/workplacelearningtoday/?p=767

Dharinya, V. S. & Jayanthi, M. K. (2013), *Effective Retrieval of Text and Media Learning Objects Using Automatic Annotation*. World Applied Sciences Journal, Volume Vol.27(1), pp. 123-129.

Greenberg, J. (2004), *Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications*. Journal of Internet Cataloguing: The International Quarterly of Digital Organization, Classification, and Access, 6(4), pp. 58-82.

Marcus-Quinn, A. & Diggins, Y. (2013), *Open Educational Resources*. Procedia-Social and Behavioral Sciences, Volume Vol.93, pp. 243-246.

Miranda, S. *et al.* (2013), *Automatic Generation of Assessment Objects and Remedial Works for MOOCs*. Proc. of ITHET 2013, Antalya, Turkey, October 10-12 2013, ISBN: 978-972-8939-88-5, pp. 175-182.

Miranda, S. & Ritrovato, P. (2014), *Automatic extraction of metadata from learning objects*. Salerno, Italy, 6th IEEE International Conference on Intelligent Networking and Collaborative Systems, INCoS 2014.

Ronsivalle, G., Carta, S. & Metus, V. (2009), *L'arte della progettazione didattica. Dall'analisi dei contenuti alla valutazione dell'efficacia*. Milano: FrancoAngeli.

Sagi, E., Kaufmann, S. & Clark, B. (2009), *Semantic density analysis: comparing*

word meaning across time and phonetic space. Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (Athens, Greece, March 31 - 31, 2009). M. Pennacchiotti, Ed. ANLP/NAACL Workshops. Association for Computational Linguistics, pp. 104-111.

Shannon, C. E. (1993), *A Mathematical Theory of Communication*. Bell System Technical Journal, July and October 1948, Volume Reprinted in Claude Elwood Shannon: Collected Papers. New York: IEEE Press.