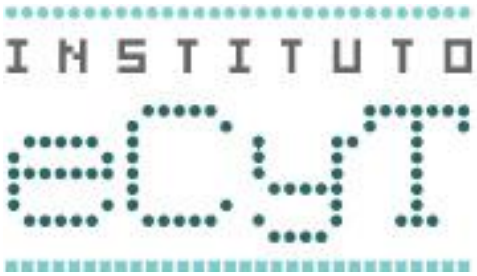


VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Coding science news (intrinsic and extrinsic features)

MIGUEL ÁNGEL QUINTANILLA,
CARLOS G. FIGUEROLA
TAMAR GROVES

Science news in Spain

- The corpus of digital news articles: problems and proposed solutions
- The selection of science and technology articles: work in progress
- Automatic coding (intrinsic and extrinsic features of science): there are reasons to be optimistic
- Preliminary indicators

The corpus of news articles

- We used the digital versions of three national newspapers: El Mundo, El País y El Público from 2002 to 2011 (El Público only from 2007).
- The reasons for our choice:
  - El Mundo and El País are the largest newspapers in Spain
  - The editors of El Público declared that science and technology would receive special attention

The corpus of news articles: using digital versions

Advantages:

- Easy to process automatically
- Permits managing large scale data
- There is a growing tendency to use the digital press

Disadvantages:

- It's a dynamic source that changes continuously
- Due to the structure of the WebPages some parts are less accessible than others
- It is difficult to identify and isolate single articles

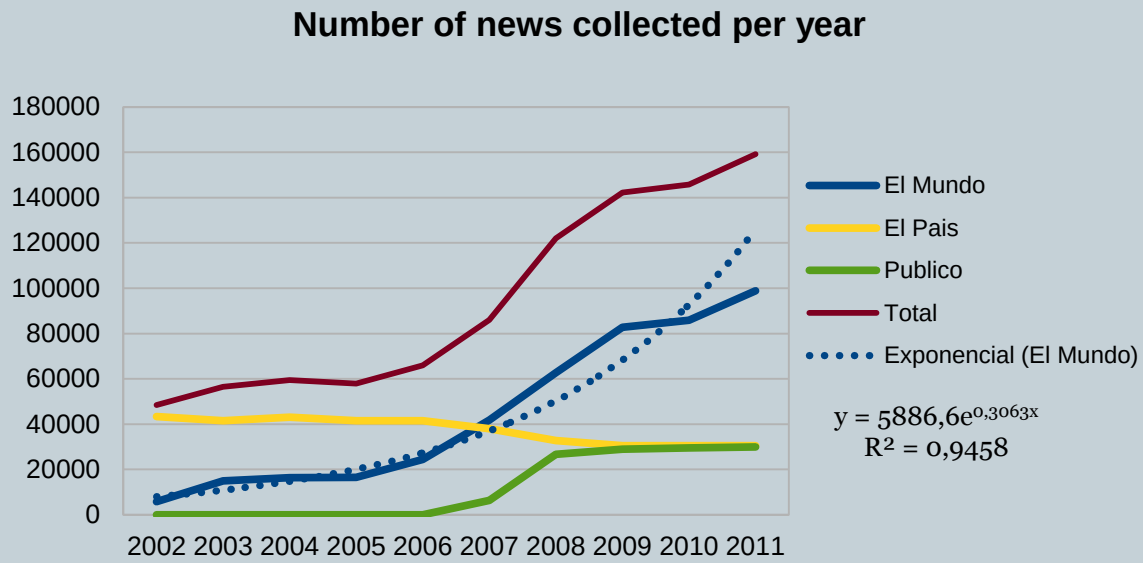
The corpus of news articles: Methodology

- A web crawler was trained to recover as much information as it was able to
- Automatic information retrieval procedures were used
- The articles were saved in HTML format
- They were identified and isolated as much as possible from other elements such as publicity, banners, menus, etc
- They were converted into plain text

The corpus of news articles: Problems: I

- There are parts of the digital versions that disappeared
- Due to changes in the structure of the WebPages it is impossible to guarantee that we gathered all the articles that have appeared on the digital versions
- The same article can be associated with two or more different URL

The corpus of news articles: Problems: II



The corpus of news articles: data

	El Mundo	El Pais	Publico	Total
2002	5714	43315		49029
2003	14966	41537		56503
2004	16384	43105		59489
2005	16434	41517		57951
2006	24449	41515		65964
2007	41724	37926	6367	86017
2008	82757	32711	26632	122100
2009	82805	30448	28976	142229
2010	85821	30484	29434	145739
2011	98918	30458	29893	159269
TOTAL	449972	373016	121302	944290

The corpus of news articles: Preliminary considerations

- We can improve the technical procedures, but automatic retrieval procedures will never be perfect as the corpus of digital news is up to a certain degree a fuzzy entity.
- While we are not dealing with exact numbers, we are able to process big data.

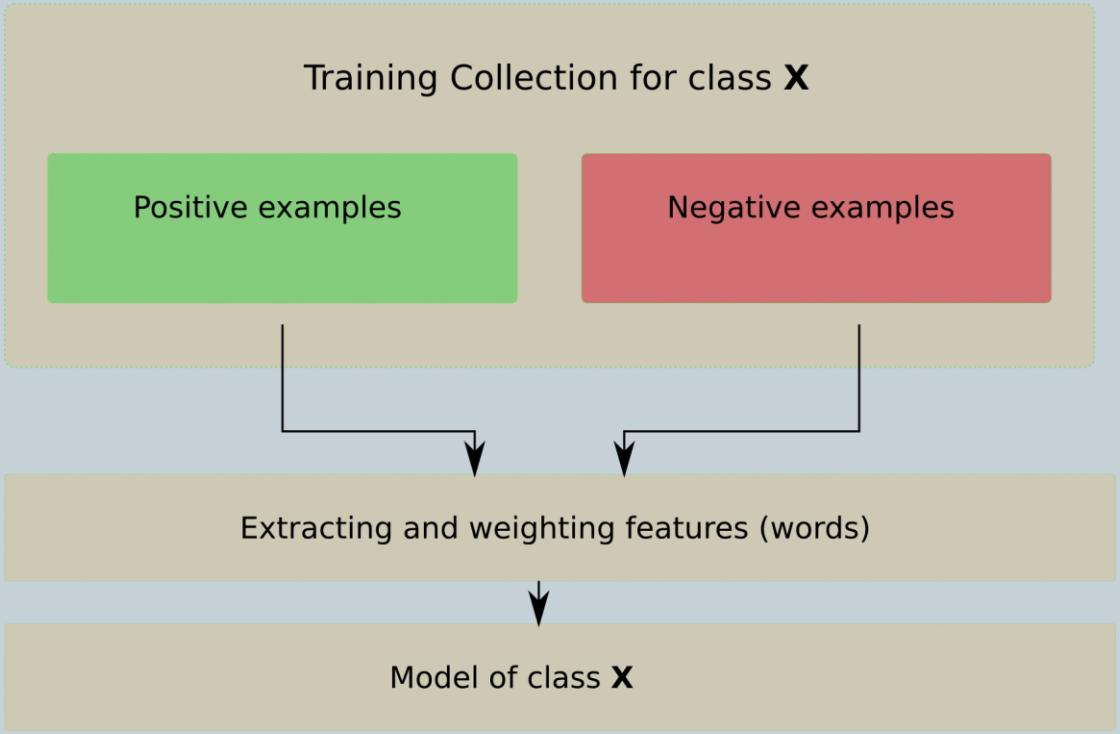
The selection of science and technology articles: Methodology

- We used an automatic supervised classifier.
- The classifier was trained to identify contents of science and technology, using the science and medicine sections of El Mundo.
- We selected a training sample of 999 articles identified as science and technology according to this procedure.
- This sample was revised and coded manually.

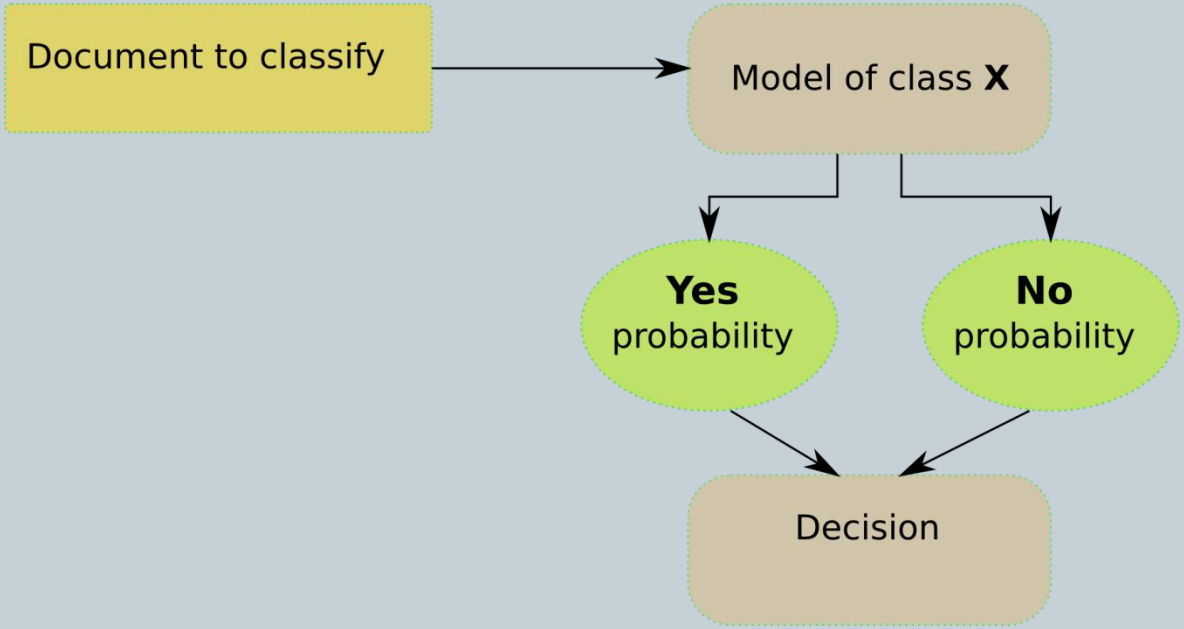
The selection of science and technology articles: Problems

- We detected a high presence of articles from El Mundo.
- In order to fix it we created a new training sample of 600 articles adjusted to the relative weight of each newspaper.
- To compensate the loss of articles from the training sample, we used an iterative active learning procedure.

The selection of science and technology articles: The training phase



The selection of science and technology articles: The classifying phase



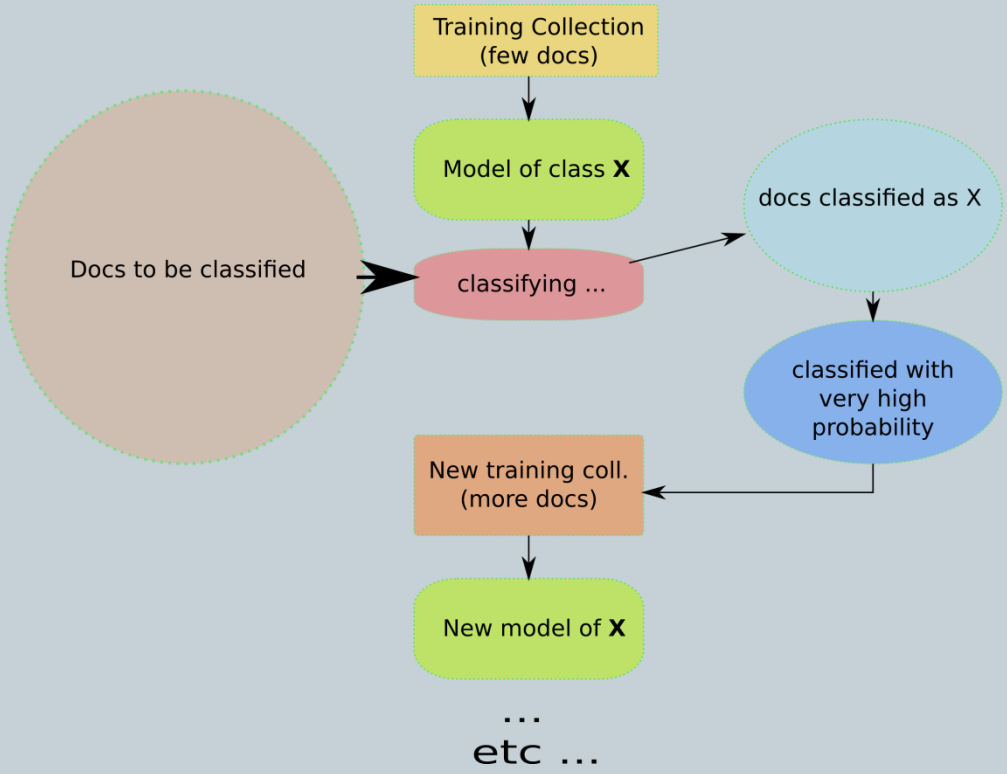
The selection of science and technology articles: The classifying phase

- The classifier we used is: Support Vector Machines (SVMs)
- We employed libsvm for python
- We opted for binary classification according to each category independently from other categories

The selection of science and technology articles: The classifying phase

- Vapnik, Vladimir N.; *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995. ISBN 0-387-98780-0
- About SVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- Guide SVM: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
- Information and Software: http://svmlight.joachims.org/

The selection of science and technology articles: The re-training phase



The selection of science and technology articles: The re-training phase

- We run a test called k fold cross-validation in order to evaluate the accuracy of the classifier (K=5).
- The results we obtained were:

C	98,6	T	97,55
CI	82,07	CE	83,77
TE	86,36	TI	77,9
CyT	98,23		
Avg	89,21		

The selection of science and technology articles: Results

Science and Technology						
	Sci orTech	Sci	Tech	Sci and Tech	Sci nor Tech	Total ST*
2002	2111	1567	758	214	68	2179
2003	2799	2138	930	269	84	2883
2004	2743	2147	789	193	131	2874
2005	3059	2353	871	165	131	3190
2006	4560	3783	955	178	113	4673
2007	6482	5419	1359	296	164	6646
2008	8248	6652	2007	411	298	8546
2009	8406	6784	1979	357	174	8580
2010	8552	6409	2589	446	140	8692
2011	10270	7112	3676	518	373	10643
TOTAL	57230	44364	15913	3047	1676	58906

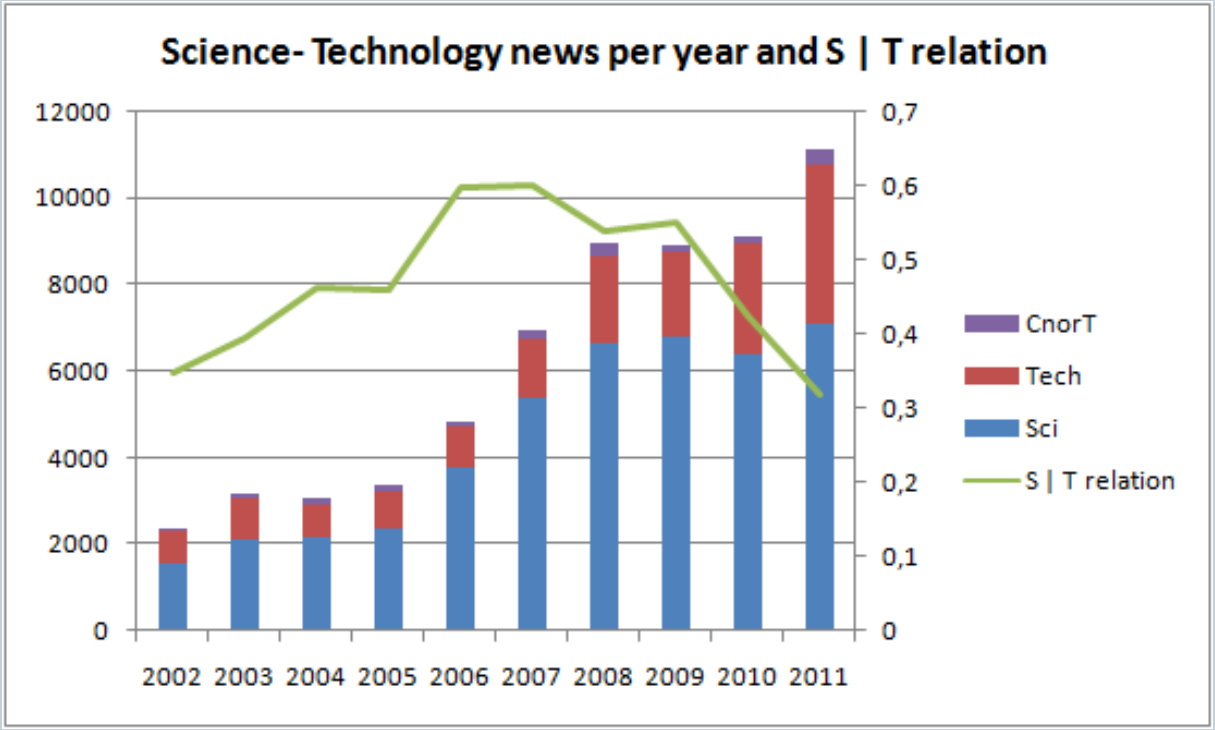
The selection of science and technology articles: Results

% of total number of articles

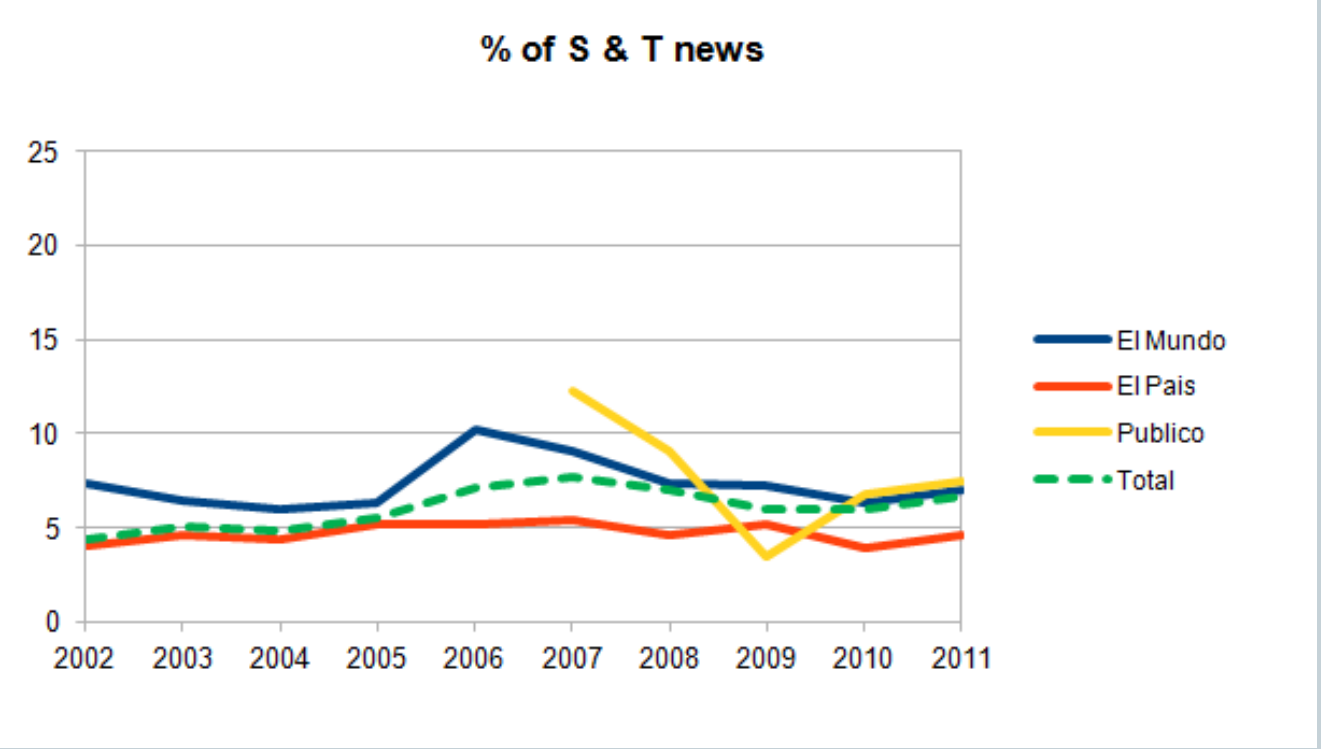
% Science and technology on total news							
	All News	Sci orTech	Sci	Tech	Sci and Tech	Sci norTech	S|T*
2002	49029	4,31	3,20	1,55	0,44	0,14	0,3480
2003	56503	4,95	3,78	1,65	0,48	0,15	0,3937
2004	59489	4,61	3,61	1,33	0,32	0,22	0,4625
2005	57951	5,28	4,06	1,50	0,28	0,23	0,4597
2006	65964	6,91	5,73	1,45	0,27	0,17	0,5969
2007	86017	7,54	6,30	1,58	0,34	0,19	0,5990
2008	122100	6,76	5,45	1,64	0,34	0,24	0,5364
2009	142229	5,91	4,77	1,39	0,25	0,12	0,5483
2010	145739	5,87	4,40	1,78	0,31	0,10	0,4245
2011	159269	6,45	4,47	2,31	0,33	0,23	0,3185
TOTAL	944290	6,06	4,70	1,69	0,32	0,18	0,4720

*S|T = (SCI-TECH) / (SCI+TECH)

The selection of science and technology articles: Results



The selection of science and technology articles: Results



The selection of science and technology articles: Preliminary considerations

- The absolute numbers are not reliable due to the problem we had with regard to El Mundo.
- As the problem affected the whole corpus the percentages can be indicative, although we will have to confirm our results once we solve the problem.

The selection of science and technology articles: Preliminary considerations

Based on the data gathered up to this point:

- The weight of Science and Technology in the corpus of news is 6 % (min 4.31 % 2002, max 7.54 % 2007).
- An indicator of the relative importance of science (from -1 to 1) has values between 0.32 in 2011 and 0.6 in 2007

Automatic coding

- The same automatic classifier was used to code the articles according to their intrinsic and extrinsic features.
- Based on the same sample of 999 articles coded manually was used in order to train the classifier.

Automatic coding: model

		SCIENCE	TECHNOLOGY
Intrinsic culture	**Representations (Representational)**	Knowledge Scientific information (IRSC)	Knowledge Technological information (IRTC)
	Practices (Operational)	Rules of the scientific method. Scientific mode of action (IOSC)	User Guide and rules of usage of technology. (IOTC)
	Values (Evaluative)	Scientific values: objectivity, precision etc (IESC)	Technological values: efficiency, reliability etc (IETC)
Extrinsic culture	**Representations (Representational)**	Images of science (ERSC)	Images of technology (ERTC)
	Practices (Operational)	Interest in science. Norms of behaviour (moral, legal etc) related to science (EOSC)	Interest in technology. Norms of behaviour (moral, legal etc) related to technology (EOTC)
	Values[1] (Evaluative)	Evaluations and attitudes toward science (EESC) (EESC +) (EESC -)	Evaluations and attitudes toward technology (EETC) (EETC+) (EETC-)

Automatic coding: Results

Intrinsic/Extrinsic Science features per year

	SI	SE	SI and SE	SI or SE	SI nor SE.	All Sci news
2002	1021	361	253	1129	438	1567
2003	1489	396	277	1608	530	2138
2004	1399	448	304	1543	604	2147
2005	1352	315	223	1444	909	2353
2006	2529	479	327	2681	1102	3783
2007	3561	737	400	3898	1521	5419
2008	3972	580	320	4232	2420	6652
2009	3670	426	228	3868	2916	6784
2010	3933	575	319	4189	2220	6409
2011	4637	575	350	4862	2250	7112
TOTAL	27563	4892	3001	29454	14910	44364

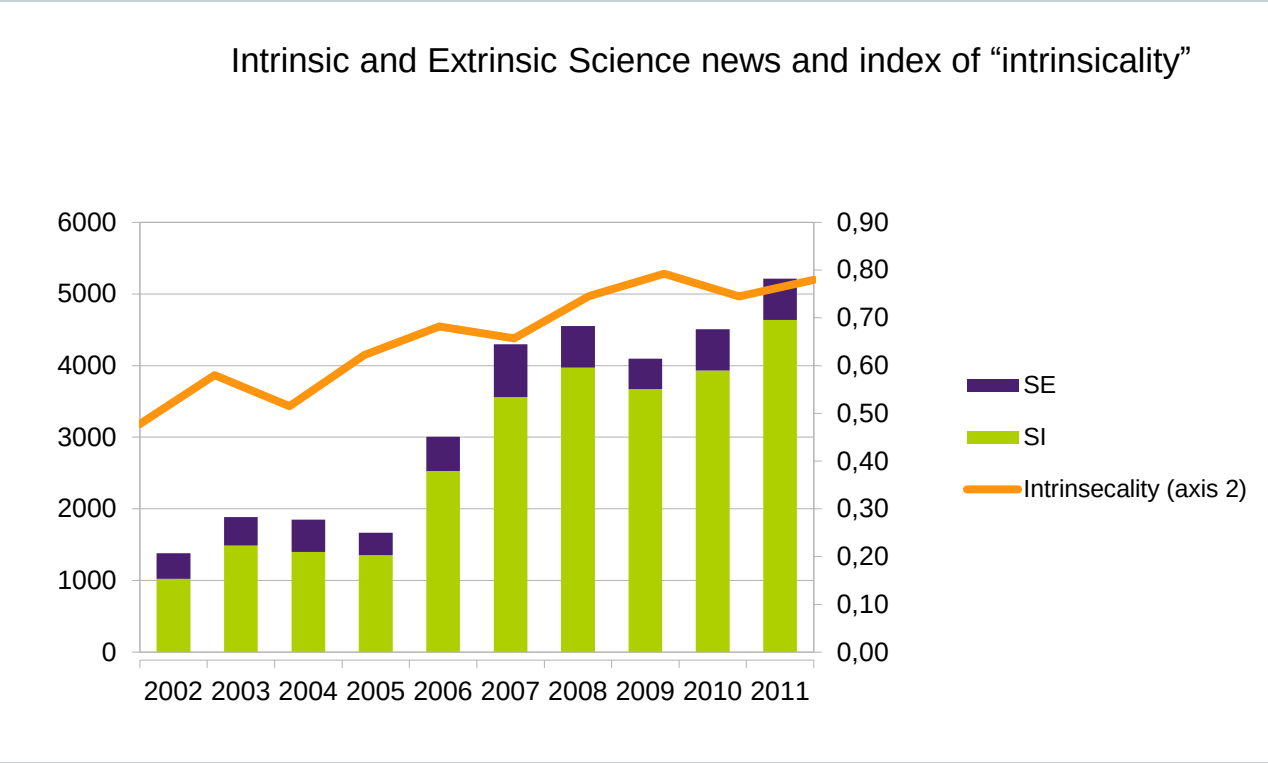
Automatic coding: Results

Intrinsic/Extrinsic per year, % of scientific articles and indicator of "intrinsicality"

	SI	SE	SI and SE	SI or SE	SI nor SE	Intrinsecality
2002	65,16	23,04	16,15	72,05	27,95	0,48
2003	69,64	18,52	12,96	75,21	24,79	0,58
2004	65,16	20,87	14,16	71,87	28,13	0,51
2005	57,46	13,39	9,48	61,37	38,63	0,62
2006	66,85	12,66	8,64	70,87	29,13	0,68
2007	65,71	13,60	7,38	71,93	28,07	0,66
2008	59,71	8,72	4,81	63,62	36,38	0,75
2009	54,10	6,28	3,36	57,02	42,98	0,79
2010	61,37	8,97	4,98	65,36	34,64	0,74
2011	65,20	8,08	4,92	68,36	31,64	0,78
Total	62,13	11,03	6,76	66,39	33,61	0,70

Intrinsicality = (SI-SE) / (SI+SE)

Automatic coding: Results



Automatic coding: Results

Intrinsic/Extrinsic Technology features per year

	TI	TE	TI and TE.	TI or TE	TI norTE.	All Tech news
2002	64	530	64	530	228	758
2003	34	628	33	629	301	930
2004	34	585	34	585	204	789
2005	42	643	42	643	228	871
2006	60	713	60	713	242	955
2007	65	943	65	943	416	1359
2008	67	1421	67	1421	586	2007
2009	146	1481	143	1484	495	1979
2010	230	1903	229	1904	685	2589
2011	249	2851	249	2851	825	3676
TOTAL	991	11698	986	11703	4210	15913

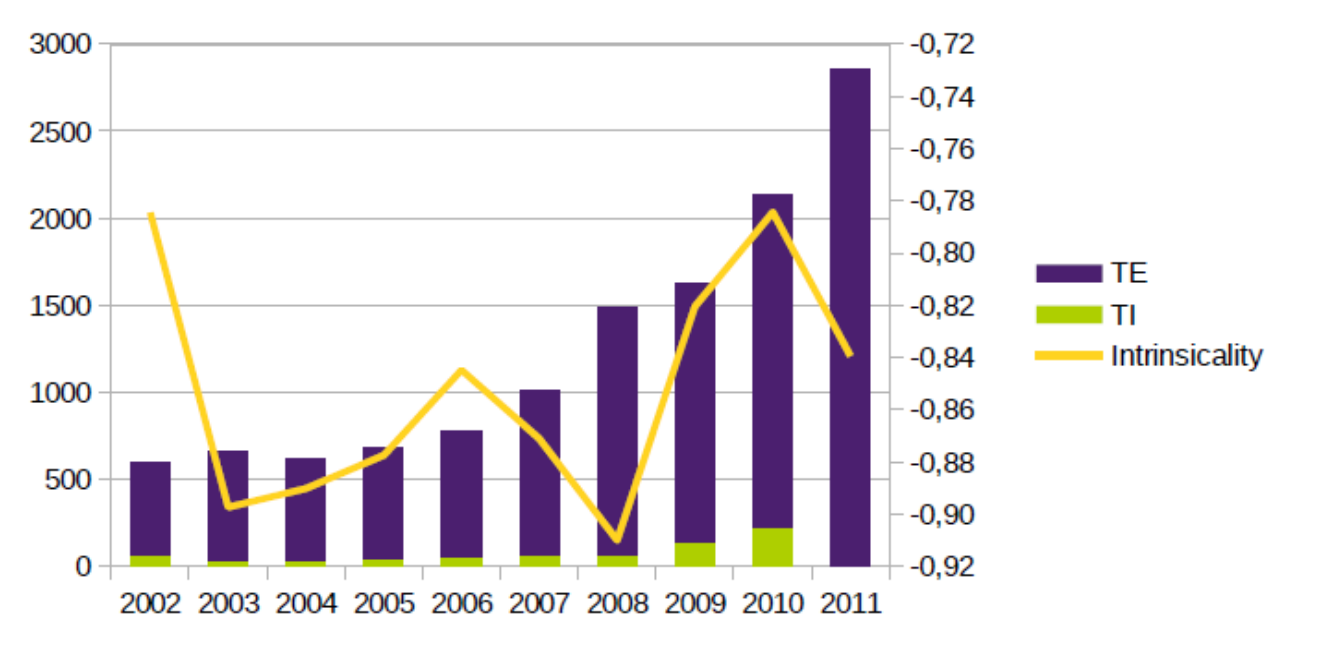
Automatic coding: Results

Intrinsic/Extrinsic per year, % of technological articles and indicator of "intrinsicality"

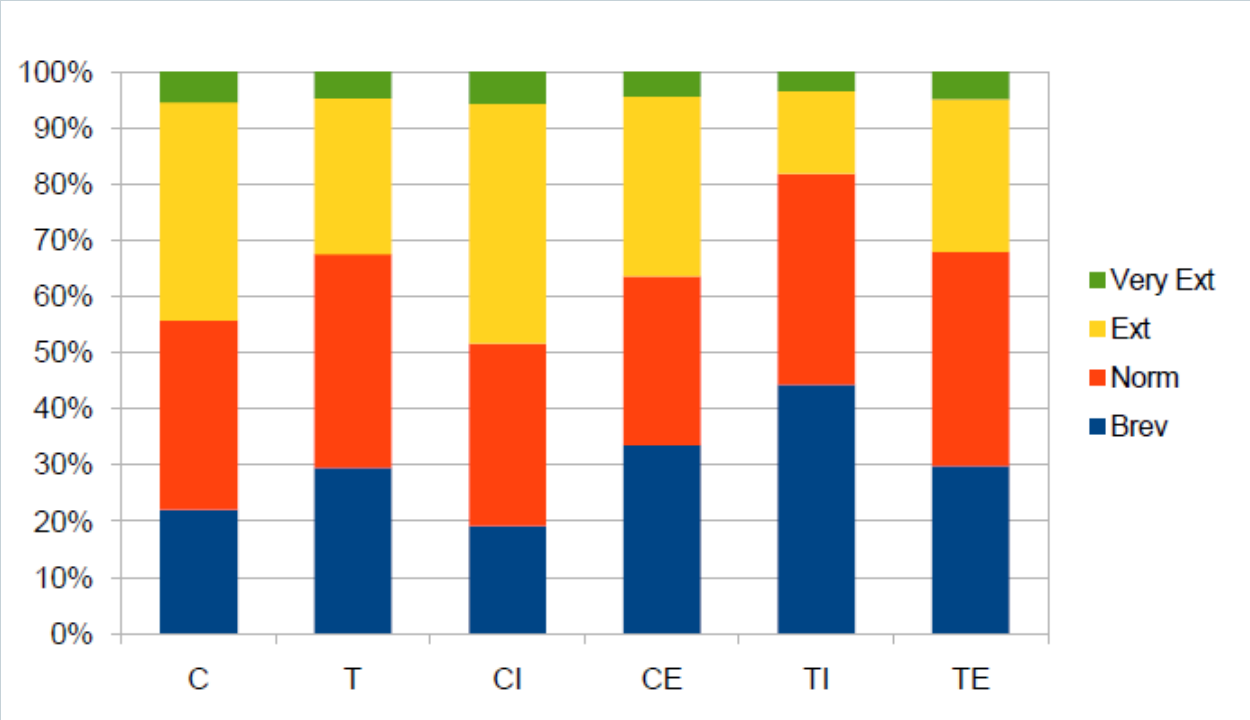
	TI	TE	TI and TE.	TI or TE	TI norTE.	Intrinsicality (secondary axis)
2002	8,44	69,92	8,44	69,92	30,08	-0,78
2003	3,66	67,53	3,55	67,63	32,37	-0,90
2004	4,31	74,14	4,31	74,14	25,86	-0,89
2005	4,82	73,82	4,82	73,82	26,18	-0,88
2006	6,28	74,66	6,28	74,66	25,34	-0,84
2007	4,78	69,39	4,78	69,39	30,61	-0,87
2008	3,34	70,80	3,34	70,80	29,20	-0,91
2009	7,38	74,84	7,23	74,99	25,01	-0,82
2010	8,88	73,50	8,85	73,54	26,46	-0,78
2011	6,77	77,56	6,77	77,56	22,44	-0,84
TOTAL	6,23	73,51	6,20	73,54	26,46	-0,84

Automatic coding: Results

Intrisic and Extrinsic technological articles and intrisicality index



The selection of science and technology articles: Results



Preliminary indicators:

Indicators	Manual coding (999)	Results for the whole corpus
S|T	0.44	0.47
SI|CE	0.67	0.70
TI|TE	-0.38	-0.84
I|EX	0.31	0.27

Final considerations:

- The automatic retrieval procedure should be improved in order to overcome the current abnormalities
- The selection of science and technology articles proved to be quiet reliable, although the automatic classification is not able to distinguish in some cases between the different categories.
- the comparison between the manual and automatic coding leads us to believe that the automatic coding can be reliable.