Vetenskapsrådet

SUBJECT CLASSIFICATION OF PUBLICATIONS IN THE ISI DATABASE BASED ON REFERENCES AND CITATIONS

VETENSKAPSRÅDETS LILLA RAPPORTSERIE 4:2011

SUBJECT CLASSIFICATION OF PUBLICATIONS IN THE ISI DATABASE BASED ON REFERENCES AND CITATIONS

Magnus Gunnarsson, Johan Fröberg, Carl Jacobsson, Staffan Karlsson

Swedish Research Council

SUBJECT CLASSIFICATION OF PUBLICATIONS IN THE ISI DATABASE BASED ON REFERENCES AND CITATIONS

VETENSKAPSRÅDET
Box 1035
101 38 Stockholm

ISBN 978-91-7307-197-0

ABSTRACT

The Thomson Reuters/ISI database of scientific publications includes a subject classification of the journal issues. One of the subject classes is Multidisciplinary Sciences, which includes such prestigious journals as Nature and Science, among others. This means that comparisons based on the subject classes in the database treat articles in Nature and Science papers separately from similar articles in more specified journals. For example, a medicine article published in Nature is not compared to other medicine articles, but rather to other articles published in multidisciplinary journals.

This paper describes the method for reclassifying papers in multidisciplinary journals based on the papers' references and citations used in the publication database at the Swedish Research Council. The method manages to reclassify more than 50% of the papers in multidisciplinary journals. Most of the papers that the method fails to classify are lowly cited.

CONTENTS

INTRODUCTION

The Thomson Reuters/ISI citation database, used in many bibliometric studies, is organised as a set of issues containing 1 or more papers[1]. All, or almost all, of the issues in the database are classified into one or more of 256 subject categories. Normally, all issues of a given journal are classified in the same way.

One of the subject categories is *Multidisciplinary Sciences*, used for journals (issues) containing papers from several scientific disciplines. It should be noted that it is the journal that is classified as multidisciplinary, not the individual papers. *Multidisciplinary sciences* thus means that the journal (issue) in question contains papers from several different disciplines, not that the papers as such necessarily concern several disciplines. Papers considered to be multidisciplinary are handled by assigning more than one subject tag to the relevant journal.

It should also be noted that several very prestigious journals are classified as Multidisciplinary Sciences, such as *Nature* and *Science*. This means that bibliometric analyses that take subject classification into account when comparing papers usually either exclude several important journals or treat papers in these journals as a separate case, to be compared with other papers in these journals, rather than with other papers in their "true" subject class. For example, a paper describing oncology research that is published in *Nature* will have to gain at least 14 citations in order to reach the field average, while if it is published in the journal O*ncology* it only has to gain 6–7 citations to reach field average.

The present work describes a method to reclassify the papers in journal issues belonging to the class Multidisciplinary Sciences, placing them in one or more of the other 255 subject classes. The reclassifications are based on the subject classes of the papers that refer to, or are referred by, the papers to be reclassified. The procedure works in two steps, where the first is a classification based on references, and the second is one based on citations[2]. The first part of the report compares the two methods when used separately, and the second part present the result obtained when both methods are combined. The combined method is used in the database at the Swedish Research Council.

[1] To be precise, the issues contain *items*, which can be of several different types, such as *article*, *review*, and *editorial note*.

[2] The terms *reference* and *citation* denotes the same thing from different perspectives. When paper A contains a reference to paper B, then it is a reference from the perspective of A and a citation from the perspective of paper B.

PART 1: COMPARISON OF REFERENCE AND CITATION CLASSIFICATIONS

A reference-based classification scheme

Although papers in journals classified as *Multidisciplinary Sciences* are not in themselves necessarily multidisciplinary, we shall refer to them here as multidisciplinary papers, in order to simplify the description. When there is a risk of confusion, a more precise terminology will be used. It should also be remembered that "multidisciplinary papers" are papers for which at least one of the subject tags is *Multidisciplinary Sciences* – there may be other subject tags as well.

Classifying multidisciplinary papers based on the subject profile of their references relies on the presupposition that a paper "intrinsically" belonging to a certain field primarily refers to other papers of that field. For example, the reference list of a medicine paper in *Nature* will be dominated by papers published in medicine journals. This presupposition and method has been used before (Glänzel et al 1999a, 1999b). We use a somewhat different algorithm here, and a comparison of these two algorithms is still to be done.

Unsurprisingly, the multidisciplinary papers in the database often refer to other multidisciplinary papers. Because of this a reference-based classification of any given paper will work better if the papers it refers to have already been classified, ceteris paribus. Further, since the database only contains papers within a limited time span, currently, 1982–2010, and since references almost always concern papers older than the paper they appear in, a classification based on references will work better on new papers than on old papers, again ceteris paribus. For example, a reference-based classification will hardly work at all for papers published in 1982, since these papers in the vast majority of cases will refer to papers which were published before 1982, and thus are not included in the database[3]. All this leads to the conclusion that a reference-based classification of multidisciplinary papers should start with the oldest papers and work its way towards the newest ones. The following example illustrates this:

> *Paper A was published in 2001 in a multidisciplinary journal, and it refers to another paper B, also published in a multidisciplinary journal, in 1998. When the year 1998 is processed, B is classified as a biology paper. When the year 2001 is processed, the classification of A is aided by the fact that its reference to B is a reference to a biology paper, rather than to a multidisciplinary paper.*

The algorithm is described in detail in appendix 2; certain details is dealt with there, such as how to handle multidisciplinary papers that also have other subject tags, and how to make sure that the number of subject tags for any individual paper does not exceed six.

Subject Determination

To determine the subject of a paper we shall look at the subjects of the papers it refers to. To simplify the description, we say that a paper *A refers to the subject S* if and only if there is a reference from *A* to another paper *B* which has been classified as *S*.

The basic principle of the method is to count the number of references for each subject referred to by a paper, and let the most commonly referred subject be the new subject of the paper. Since the ISI practice allows up to six subjects per article (journal issue), we will in fact let the most commonly referred subjects be the new subjects of the paper.

[3] Our database starts with 1982

Let's pretend four papers have the following distributions of subject references:

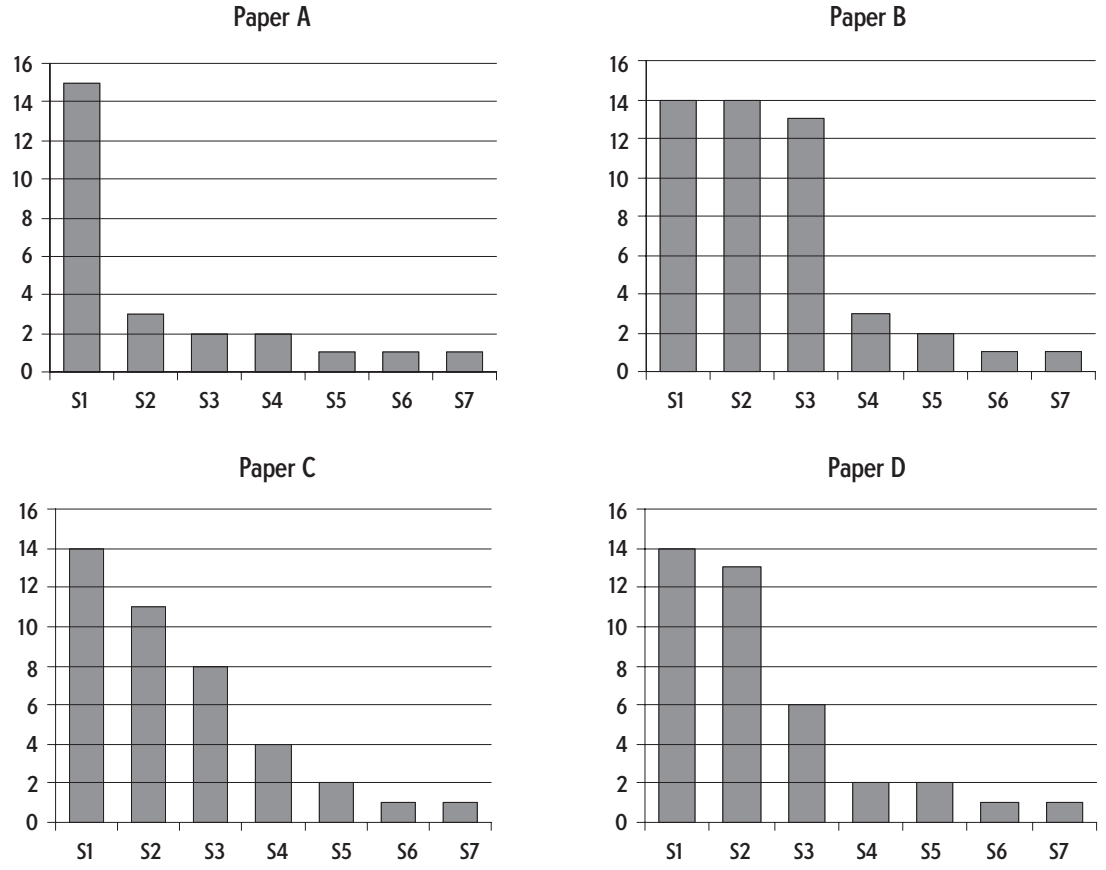

Figure 1: Distribution of subjects for the references of four hypothetical papers. For paper A, subject S1 has 15 references; subject S2 has 3 references; etc.

For each of the four hypothetical papers in Figure 1, which of the subject tags S1-S7 should be assigned to the paper? For paper A, S1 seems like the obvious choice, and for paper B S1, S2 and S3 seem equally obvious. We have used the following, somewhat arbitrary principle:

1 The most common subject is assigned to the paper.
2 Assuming that the referred subjects of a paper are ordered as in figure 1 above, a subject is assigned to a paper if the number of references to that subject is 60% or more of the number of references to the preceding subject.

For example, if subject S1 has 14 references and subject S2 has 10 references, then S2 is included, since 10 is more than 60% of 14 (cf paper C in figure 1).

In order for this principle to be reasonable, there must be a most common subject. In order to handle cases where the subjects are many and evenly distributed, we add the following rule:

3 If the six most common subjects do not total 60% of all references, then the paper is considered truly multidisciplinary, and no reclassification is done.

The reason for drawing the line at the six most common subjects comes from the fact that the ISI database uses at most six subject tags per journal issue. For the same reason, a maximum of six subject tags are assigned to each paper:

4 Only the six most commonly referred subjects are candidates for classification.

If should be noted that the number of subjects referenced by a paper is not fractionalised, i.e. a single reference to a paper with two subjects A and B is counted as two references, one to subject A and one to subject B.

Some papers have very few references. Since we do not think that a paper with very few references is a good candidate for being classified based on its references, we require a minimum of three references:

5 Only papers with three or more references are candidates for classification.

The limit of three references is rather arbitrarily chosen. It seemed reasonable to us.

The precise algorithm, including details about how to handle papers for which multidisciplinary papers is only one of several subject tags, is described in appendix 1.

A citation-based classification scheme

Classifying papers according to their citations can be done in very much the same way as classifying them according to their references, the main difference being that the process should start with the newest papers and work its way back to the oldest ones, since papers almost always are cited by younger papers. Another and perhaps more subtle difference is that the meaning of the classification changes somewhat: the citations of a paper will reflect the subject(s) for which the paper has had the greatest impact, while the references of a paper will reflect the subjects which the paper has been influenced by.

Results of the first study

Table 1: Results of reference- and citation-based classifications.

	All document types		Articles & reviews	
	Abs.	**Rel.**	**Abs.**	**Rel.**
Number of multidisciplinary papers initially:	523 901	100%	364 871	100%
Number of multidisciplinary papers after a reference-based classification:	320 491	61%	175 268	48%
Number of multidisciplinary papers after a citation-based classification:	330 086	63%	190 200	52%
Number of multidisciplinary papers after a reference-based classification that is not multidisciplinary papers after a citation-based classification:	50 515	10%	38 819	11%
Number of multidisciplinary papers after a citation-based classification that is not multidisciplinary papers after a reference-based classification:	60 110	11%	53 751	15%

Table 1 shows the results of reference- and citation-based classifications of multidisciplinary papers. The two types of classifications manage to reclassify approximately the same number of papers, just below 40% of all multidisciplinary papers and 50 % of the articles and reviews.

Table 2 below shows the preconditions for reference- and citation-based classifications.

Table 2: Preconditions for reference- and citation-based classifications.

	All document types		Articles & reviews	
	Abs.	Rel.	Abs.	Rel.
Number of multidisciplinary papers:	523 901	100%	364 871	100%
Number of multidisciplinary papers without references:	150 737	29%	59 752	16%
Number of multidisciplinary papers without citations:	257 712	49%	136 079	37%
Number of multidisciplinary papers for which all references lead to other multidisciplinary papers:	25 513	5%	17 060	5%
Number of multidisciplinary papers for which at least 50% of the references lead to other multidisciplinary papers:	34 356	7%	23 430	6%
Number of multidisciplinary papers for which all citations come from other multidisciplinary papers:	13 371	3%	9 714	3%
Number of multidisciplinary papers for which at least 50% of the citations lead to other multidisciplinary papers:	15 219	3%	11 239	3%

Taking the preconditions into account, the citation-based classification method manages to classify more papers than the reference-based one does: About 80% of the papers that have citations were classified by the citation-based method, while only 61% of the papers that have references were classified by the reference-based method. This is largely explained by the high average number of citations per paper, as compared to the average number of references per paper.

Agreement between the reference- and citation-based classifications

A simple measure of the agreement between the reference and citation classifications is the average relative overlap, i.e. the ratio of the number of subjects for one classification and the number of subjects for both classifications.

For example, if the reference classification assigns two subjects A and B for a given paper, and the citation classification assigns the same two subjects A and B to that paper, then the overlap is 2/2=1 for both the reference and the citation classification. If the reference classification assigns 10 subjects A,B,C,D,E,F,H,I,J and K to a given paper, and the citation classification assigns 2 subjects K and L to that paper, then the relative overlap is 10/11=0.91 for the reference classification and 1/11=0.09 for the citation classification.

Using formal notation this can be described the following way:

$$V_{ref}(i) = \frac{N_{ref,i}}{N_{tot,i}}$$

where $V_{ref}(i)$ is the relative overlap for the reference classification for paper i, $N_{ref,i}$ is the number of subjects assigned to i by the reference classification, and $N_{tot,i}$ is the number of subjects assigned to the paper i by the reference classification together with the citation classification; doubles not counted. The relative overlap for the citation classification for the paper i, $V_{cit}(i)$ is calculated correspondingly.

A combined measure for the reference and citation overlaps is constructed by simply adding the two quotients. Subtracting 1 gives the measure a more appealing span of 0-1:

$$Combined\ relative\ overlap = \frac{N_{ref,i}}{N_{tot,i}} + \frac{N_{cit,i}}{N_{tot,i}} - 1$$

The average relative overlap for the reference classification was 0.831 and for the citation classification it was 0.843. Figure 2 shows the frequency of different degrees of combined relative overlap.

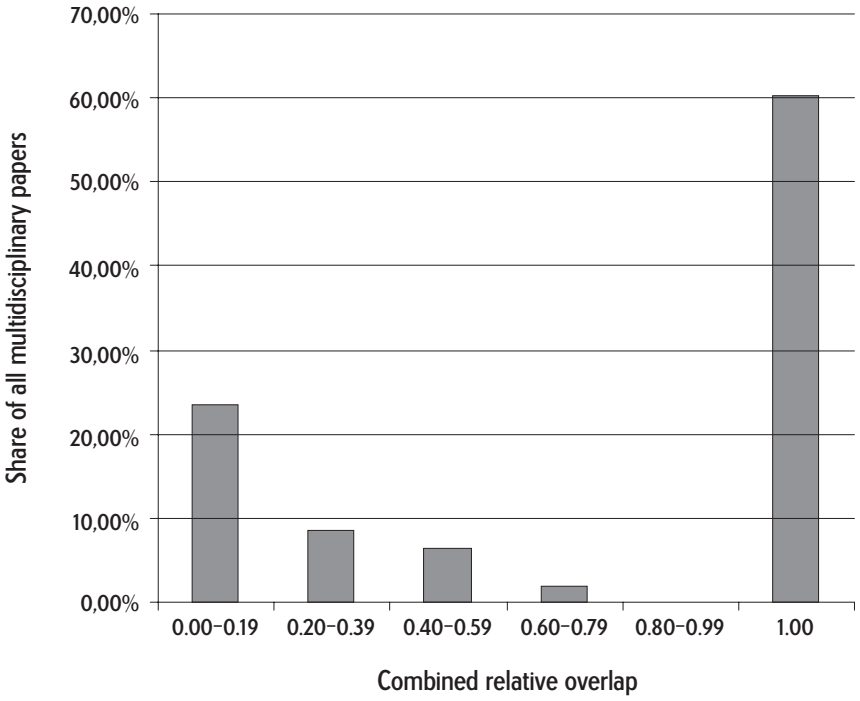

Figure 2: Frequency of degree of overlap between reference based and citation based classification.

The diagram above shows that almost 25% of the mulitdisciplinary papers had an overlap between 0 (inclusive) and 0.2 (non-inclusive), about 8% of the papers had an overlap between 0.2 and 0.4, etc. One interpretation of the diagram is that when the two methods of classifications disagree on which subject tags to assign to a paper, then they disagree distinctively. A small, but existing, degree of disagreement is uncommon.

The agreement between the reference and citation classifications can also be studied by looking at the cases when the two classifications do not agree at all, i.e. when there is no subject in the reference classification that also appears in the citation classification, for a given paper. We call this special case *maximal disagreement*. Different types of maximal disagreements can be identified based on the number of subjects in the reference and citation classifications: type 1-1 means that the reference and citation classifications have one subject each[4], type 4-1 means that the reference classification has four subjects and the citation classification has one, etc.

[4] I.e. one subject each that are different from each other, since this is a type of maximal disagreement.

Figure 3 below shows how common different cases of disagreements are.

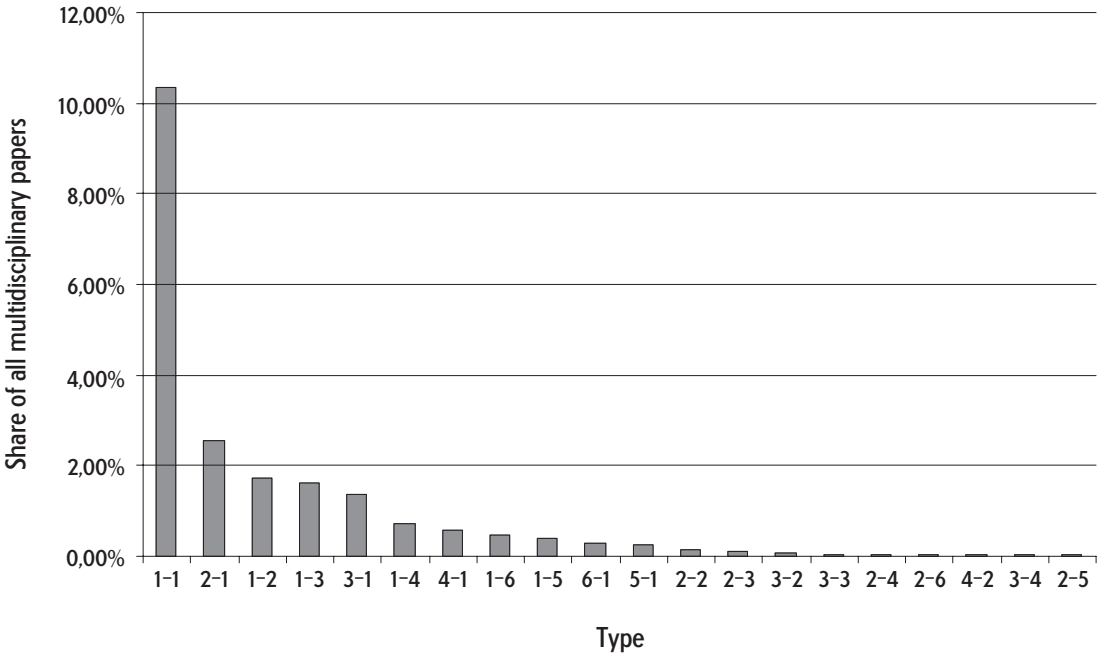

Figure 3: Frequencies for maximal disagreement types. Only the 20 most common types are shown.

Figure 3 shows that maximal disagreement usually means that the reference classification suggests one subject and the citation classification suggests one other subject.

Subjects

The subjects that the reclassified publications end up with are shown in figure 4. The largest number is moved to *Biochemistry & Molecular Biology*; 15 % of the initially multidisciplinary publications are reclassified to this subject. The second largest group after reclassification is *Neurosciences,* constituting 4.3 % of the initially multidisciplinary group. After reclassification, 7.4 % of *Biochemistry & Molecular Biology* consist of initially multidisciplinary publications. Other subject fields that get relatively large increase in volume due to the reclassification are *Paleontology* (6.6 %) and *Geochemistry & Geophysics* (5.4 %). (N.B. an initially multidisciplinary publication may end up with up to six subjects after reclassification). The total size (articles and reviews) of the twenty largest subject fields after reclassification are shown in Figure 4C.

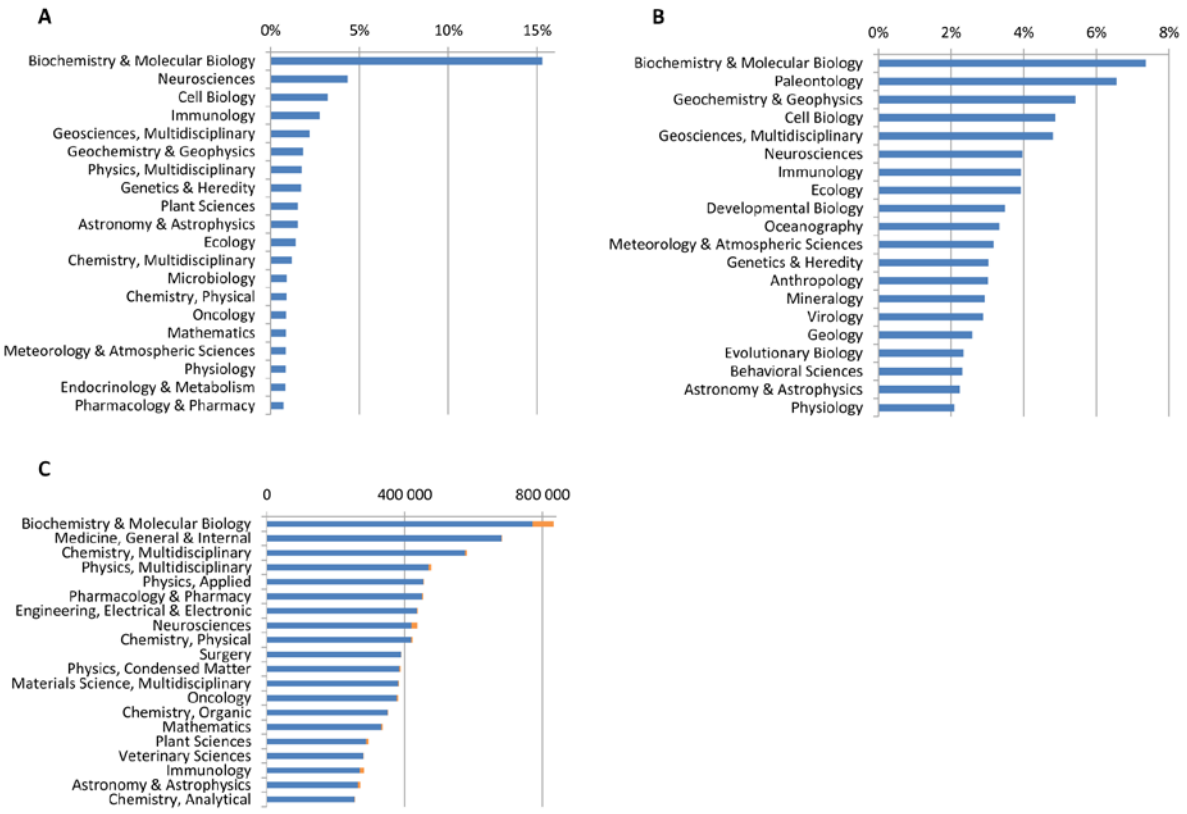

Figure 4. Subject distribution among reclassified publications. A) The twenty subject fields into which most of the initially multidisciplinary publications are reclassified. B) Contribution of reclassified, initially multidisciplinary, publications to the size of the field after reclassification. The twenty fields for which the reclassification makes the largest relative change are included. C) Size of the twenty largest subject fields after reclassification. Reclassified publications are shown in yellow. Based on articles and reviews only.

The figure shows that multidisciplinary papers usually are put in the category *Biochemistry & Molecular Biology* by the reference-based as well as the citation-based classification. The second and third most common subjetcs are *Neurosciences* and *Cell Biology*. This is possibly due to the fact that some large multidisciplinary journals are orientied towards these fields, like *Nature* and *Science*.

The citation-based and reference-based classifications display a considerable similarity, although there are certain differences as well. One example is that Immunology is more common than both *Information Science & Library Science* and *Geosciences, Multidisciplinary* in the reference-based classification, but less common than these in the citation-based classification.

The subject distribution among the papers with maximal disagreement is somewhat more extreme than that of the reference-based classification, but on the whole it follows the same pattern. The conclusion is that disagreement between the citation-based and reference-based classification is probably fairly evenly distributed among the subjects.

PART 2: COMBINING REFERENCE- AND CITATION-BASED CLASSIFICATIONS

A reference-based classification scheme could be combined with a citation-based one, and the second study concerns just that. Certain technicalities have to be dealt with, however.

First, the combination could be done in more than one way. A straight-forward method would be to simply merge the subject tags for each paper from the reference-based classification with the subject tags from the citation-based classification. A paper that is given the subjects A and B by the reference-based classification and the subjects B and C by the citation-based classification would then end up with the subject tags A, B and C. However, this could result in more than six subject tags for a single paper, which would break the ISI standard, and we thus need a way to prioritise between the subject tags.

The combination model we propose here is based on the assumption that the reference-based classification is more correct than the citation-based one, and thus the reference-based classification should be given priority. The argument is that the references reflect all the work that the paper builds on, while the citations only reflect the parts of the paper that have influenced subsequent research.

The following general method was used for combining reference and citation information to classify multidisciplinary papers:

1. Create a list L_1 with all papers in the database, along with their respective issue-based subject tags.
2. Create a list L_2 with all multidisciplinary papers in L_1 and their reference count *R*.
3. Go through all papers listed in L_2 and replace their multidisciplinary issue-based subject tags in L1 with new subject tags based on citations. Start with the youngest papers and work your way down, one year at a time, to the oldest papers.
4. Go through all papers in L_2. If *R* is at least 3, then replace the multidisciplinary issue-based subject tag with new subject tags based on references. Start with the oldest papers and work your way up, one year at a time, to the youngest papers.
5. Go through all papers in L_2. If *R* is less than 3, then replace the multidisciplinary issue-based subject tag with new subject tags based on references and citations. Do this only if the total number of references *and* citations for each paper is at least 3, and otherwise leave the paper as it is. Start with the oldest papers and work your way up, one year at a time, to the youngest papers.

The algorithm is described in detail in appendix 3; certain details is dealt with there, such as how to handle multidisciplinary papers that also have other subject tags, and how to make sure that the number of subject tags for any individual paper does not exceed six.

Some things should be noted about this algorithm. Firstly, since the reference-based classification is made after the citation-based one, the reference-based classification overrides any classification based on citations. The following example illustrates this:

> *Paper α has been published in a multidisciplinary journal, and in step 1 it is placed in list A with the subject tag 'Multidisciplinary Sciences'. In step 2 it is placed in list B. In step 3 its citations are analysed and it is given three subject tags in list A: 'Biology,''Plant Sciences' and 'Oncology'. These three subject tags replace the previous tag 'Multidisciplinary Sciences' for paper α in A. In step 4 its references are analysed (it has more than four references) and it is given the subject tags 'Biology' and 'Microbiology'. These two subject tags replace the previous three subject tags for paper α in A.*

Further, the citation-based analysis aids the reference-based analysis in determining the subject of multidisciplinary papers, since references to papers that originally were classified as multidisciplinary in many cases have been given a subject tag by the citation-based classification.

The third thing to note is that highly cited papers with few (i.e. less than three; see appendix 2) references will, in practice, be classified based on their citations, since the citations will outnumber the references greatly.

The fourth and final thing to note is that, as in the first study, the limit of three references in steps 4 and 5 is rather arbitrarily chosen. It seemed reasonable to us.

Results of the second study

Table 3: Combining a citation-based classification with a reference-based classification.

	All document types	Articles & reviews
Number of multidisciplinary papers with neither references nor citations:	128 167	50 067
Number of multidisciplinary papers with fewer than 3 references/citations:	204 910	78 219
Number of multidisciplinary papers after a citation-based classification followed by a reference-based classification	261 139	130 231
Average number of citations for multidisciplinary papers, initially (5-year citation window):	13.59	18.77
Average number of citations for the multidisciplinary papers that remains after a citation-based classification followed by a reference-based classification (5-year citation window)	0.39	0.52

Table 3 shows that combining a reference-based classification with a citation-based one is indeed a way to improve the classification, at least in terms of number of classified papers. The combined method managed to classify just above 50%. This should be compared to the results given in Table 1, which showed that the reference-based classification managed to classify 39% of the multidisciplinary papers, and the citation-based classification managed 37%.

Table 3 also shows that the papers that the combined method does not manage to classify in general are very lowly cited.

As mentioned above, journal issues in the ISI database may have more than one subject tag. This means that some of the papers tagged as multidisciplinary have other, non-multidisciplinary subject tags as well. In order to better understand the effects of the reclassification methods it may be helpful to see how many of the papers with only a multidisciplinary subject tag – referred to as *strictly multidisciplinary* here – are given other subject tags with the classification methods described here.

Table 4: Results for strictly multidisciplinary papers.

	All document types		Articles & reviews	
	Abs.	Rel.	Abs.	Rel.
I. Number of strictly multidisciplinary papers initially:	477 023	–	334 022	–
II. Number of strictly multidisciplinary papers after a reference-based classification:	273 729	57% of I	146 261	44% of I
III. Number of strictly multidisciplinary papers after a citation-based classification:	283 589	59% of I	161 090	48% of I
IV. Number of strictly multidisciplinary papers after a reference-based classification that are not strictly multidisciplinary papers after a citation-based classification:	43 258	16% of II	32 135	22% of II
V. Number of strictly multidisciplinary papers after a citation-based classification that are not strictly multidisciplinary papers after a reference-based classification:	53 118	19% of III	46 964	29% of III
VI. Number of strictly multidisciplinary papers after a citation-based classification followed by a reference-based classification:	222 494	47% of I	108 331	32% of I

The table above shows that the classification methods described here are even more successful on strictly multidisciplinary papers than on multidisciplinary papers in general. 53% of the strictly multidisciplinary papers were classified with the combined method, while only 50% of the multidisciplinary papers in general were classified with the same method.

CONCLUSIONS

We have described a method for reclassification of publications in journals belonging to the subject category *Multidisciplinary Sciences*. The method is based on an analysis of the references and citations of each publication, and it is able to classify a large share of the publications in question. The publications for which the method does not find any other subject category but *Multidisciplinary Sciences* are in the majority of cases lowly cited. The reference- and citation-based classifications are to a large extent in agreement with each other.

BIBLIOGRAPHY

Glänzel, W.; Schubert, A. & Czerwon, H.-J. (1999a) An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. I *Scientometrics*, Vol. 44, No. 3, pp. 427-439.

Glänzel, W.; Schubert, A.; Schoepflin, U. & Czerwon, H.-J. (1999b) An item-by-item subject classification of papers published in journals covered by the SSCI database using reference analysis. I *Scientometrics*, Vol. 46, No. 3, pp. 431-441.

APPENDIX 1: PRECISE ALGORITHM FOR DETERMINING THE SUBJECT OF A PAPER BASED ON ITS REFERENCES

M is a constant larger than 1.

1. The total number of subjects referenced by the paper in question is calculated, f_{tot}, and also the subject distribution of these references. Thus f_1 is the number of references for the subject with the largest number of references, subject 1; f_2 is the number of references for the subject with the second largest number of references, subject 2; etc. until f_6.
2. If $f_{tot} < (f_1+f_2+f_3+f_4+f_5+f_6) \bullet M$ the paper is considered *truly multidisciplinary*, its subject tags are left unchanged, and the subject determination is complete. If not, the subject tag of subject 1 is added to the list of subject tags of the paper (if the subject tag is not already there), and the determination continues.
3. If $f_1 > f_2 \bullet M$ or if the number of subject tags is 6, then the subject determination is complete. If not, the subject tag of subject 2 is added to the list of subject tags of the paper (if the subject tag is not already there), and the subject determination continues.
4. If $f_2 > f_3 \bullet M$ or if the number of subject tags is 6, then the subject determination is complete. If not, the subject tag of subject 3 is added to the list of subject tags of the paper (if the subject tag is not already there), and the subject determination continues.
5. If $f_3 > f_4 \bullet M$ or if the number of subject tags is 6, then the subject determination is complete. If not, the subject tag of subject 4 is added to the list of subject tags of the paper (if the subject tag is not already there), and the subject determination continues.
6. If $f_4 > f_5 \bullet M$ or if the number of subject tags is 6, then the subject determination is complete. If not, the subject tag of subject 5 is added to the list of subject tags of the paper (if the subject tag is not already there), and the subject determination continues.
7. If $f_5 > f_6 \bullet M$ or if the number of subject tags is 6, then the subject determination is complete. If not, the subject tag of subject 6 is added to the list of subject tags of the paper (if the subject tag is not already there).

APPENDIX 2: PRECISE ALGORITHM FOR THE REFERENCE-BASED CLASSIFICATION

1. Create a list L_1 with all papers in the database, along with their subject tags marked with priority 0.
2. Create a list L_2 with all multidisciplinary papers in Lland their reference count R.
3. Go through all papers listed in L_2. Start with the oldest papers and work your way up, one year at a time, to the youngest papers. For each paper, do as follows:
   - 3.1 If there are less than three references in this paper, then leave it as it is and continue with the next paper. Otherwise, remove the multidisciplinary subject tag in L_1.
   - 3.2 Add new subject tags based on references. Give these new subject tags priorities according to their order in the subject determination, so that subject 1 has priority 1, subject 2 has priority 2, etc.
   - 3.3 If there are any subject tags with priority > 0 that already exist with priority 0, then remove the tags with priority > 0.
   - 3.4 If there are more than 6 subject tags, the remove the ones with the highest priority number until only 6 subject tags remain.

APPENDIX 3: PRECISE ALGORITHM FOR THE COMBINED CLASSIFICATION

1. Create a list L_1 with all papers in the database, along with their subject tags marked with priority 0.
2. Create a list L_2 with all multidisciplinary papers in L_1.
3. Go through all papers listed in L_2. Start with the youngest papers and work your way down, one year at a time, to the oldest papers. For each paper, do as follows:
   - 3.1 If there are less than three citations to this paper, then leave it as it is and continue with the next paper. Otherwise, remove the multidisciplinary subject tag in L_1.
   - 3.2 Add new subject tags based on citations. Give these new subject tags priorities according to their order in the subject determination, so that subject 1 has priority 1, subject 2 has priority 2, etc.
   - 3.3 If there are any subject tags with priority > 0 that already exist with priority 0, then remove the tags with priority > 0.
   - 3.4 If there are more than 6 subject tags, the remove the ones with the highest priority number until only 6 subject tags remain.
4. Go through all papers listed in L_2 again. Start with the oldest papers and work your way up, one year at a time, to the youngest papers. For each paper, do as follows:
   - 4.1 If there are less than three references in this paper, then mark the paper, and continue with the next paper. Otherwise, remove the multidisciplinary subject tag in L_1.
   - 4.2 Add new subject tags based on references. Give these new subject tags priorities according to their order in the subject determination, so that subject 1 has priority 1, subject 2 has priority 2, etc.
   - 4.3 If there are any subject tags with priority > 0 that already exist with priority 0, then remove the tags with priority > 0.
   - 4.4 If there are more than 6 subject tags, the remove the ones with the highest priority number until only 6 subject tags remain.
5. Go through all papers in L_2 that were marked in step 4.1. Start with the youngest papers and work your way down, one year at a time, to the oldest papers. For each paper, do as follows:
   - 5.1 If the number of references in this paper plus the number of citations to this paper is less than 3, then leave the paper as it is and continue with the next one. Otherwise, remove the remove the multidisciplinary subject tag in L_1.
   - 5.2 Add new subject tags based on references and citations. Give these new subject tags priorities according to their order in the subject determination, so that subject 1 has priority 1, subject 2 has priority 2, etc.
   - 5.3 If there are any subject tags with priority > 0 that already exist with priority 0, then remove the tags with priority > 0.
   - 5.4 If there are more than 6 subject tags, the remove the ones with the highest priority number until only 6 subject tags remain.