Combining Evaluation Metrics with a Unanimous Improvement Ratio and its Application to the Web People Search Clustering Task

Enrique Amigo
UNED NLP & IR group
Madrid, Spain
enrique@lsi.uned.es

Javier Artiles
UNED NLP & IR group
Madrid, Spain
javart@bec.uned.es

Julio Gonzalo
UNED NLP & IR group
Madrid, Spain
julio@lsi.uned.es

ABSTRACT

This paper presents the *Unanimous Improvement Ratio* (UIR), a measure that allows to compare systems using two evaluation metrics without dependencies on relative metric weights. For clustering tasks, this kind of measure becomes necessary given the trade-off between precision and recall oriented metrics (e.g. Purity and Inverse Purity) which usually depends on a clustering threshold parameter stated in the algorithm. Our empirical results show that (1) UIR rewards system improvements that are robusts regarding weighting schemes in evaluation metrics, (2) UIR reflects improvement ranges and (3) although it is a non parametric measure, it is sensitive enough for detecting most robust system improvements. The application of UIR to the second Web People Search evaluation campaign (WePS-2) shows that UIR is able to complement successfully the results offered by a conventional metric combination approach (such as Van Rijsbergen's F measure).

General Terms

H.3.5 [INFORMATION STORAGE AND RETRIEVAL]: Online Information Systems, Web-based Service; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; I.2.7 [Artificial Intelligence]: Natural Language Processing

Keywords

Evaluation, combining metrics, web people Search, text clustering

1. INTRODUCTION

Clustering consists of grouping items according to their similarity to each other, and it has applications in a wide range of artificial intelligent problems. In particular, in the context of textual information access, clustering algorithms are employed for information retrieval (clustering text documents according to their content similarity), document summarization (grouping pieces of text in order to detect redundant information), topic tracking or opinion mining (e.g. grouping opinions about a specific topic), etc.

In such scenarios, clustering distributions produced by systems are usually evaluated extrinsically, i.e., according to their similarity to a manually produced grouping. There exists a wide set of metrics to measure this similarity, but all of them are grounded on two dimensions: (i) to what extent items in the same cluster also belong to the same group in the gold standard; and (ii) to what extent items in different clusters also belong to different groups in the gold standard. A wide set of extrinsic metrics has been proposed: Entropy and class entropy [9, 6], Purity and Inverse Purity [11], Bcubed Precision and Recall metrics [5], metrics based on counting pairs [7, 8], etc.[1] In order to evaluate systems, metrics are usually combined according to Van Rijsbergen's F function [10]:

$$F_\alpha(A, B) = A * B/(\alpha * R + (1 - \alpha) * P)$$

where the α parameter allows to assign a relative weight to each metric. After stating the α value, the system improvements according to F are checked by means of statistical significant tests over test cases.

Our research goals are:

1. Verify to what extent the clustering evaluation results can be biased by the assigned metric weighting (i.e. the value for α), even when detected differences are statistically significant.

2. Introduce a measure to quantify improvements without dependencies from metric weighting (the *Unanimous Improvement Ratio*).

In Section 2 we discuss how metric weights can bias the results of an evaluation. In Section 3 we introduce our proposal, and in Section 4 we test it with some empirical studies. Finally, in Section 5 we present a use case (the application of our methodology to the results of the Second Web People Search Evaluation Campaign) and end with some conclusions in Section 6.

2. THE EFFECTS OF METRIC WEIGHTING IN CLUSTERING TASKS

Although there is an implicit consensus among researchers that the *F measure* [10] is the best way of combining evaluation metric pairs, F requires to assign relative weights to the individual metrics involved. For some tasks this requirement is not a problem, given that (i) metrics are often

[1]See [2] for a detailed overview.

WWW2009, April 20-24, 2009, Madrid, Spain.
.

correlated and (ii) the mathematical properties of F ensure a certain robustness across different parametrizations. Therefore, sometimes the metric weights have only a minor impact of the system ranking produced by F.

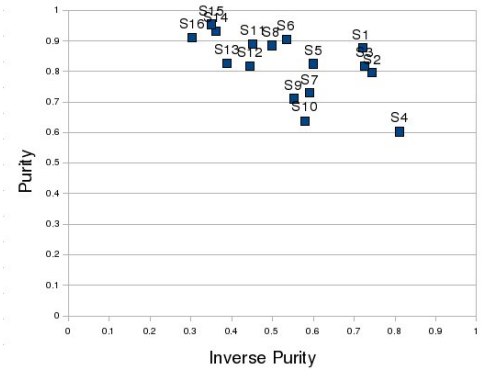

Figure 1: Evaluation results for clustering systems in WEPS 2007

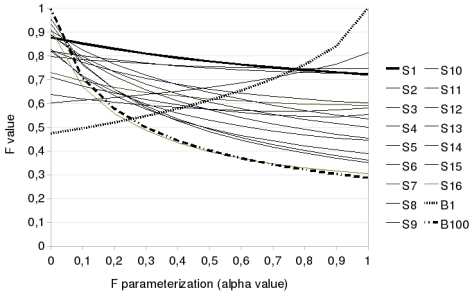

Figure 2: System evaluation results for several F parametrizations

Clustering, however, is very sensitive to the parametrization in metric combining functions. In order to obtain empirical evidence, we will analyse in this article the evaluation results for the First Web People Search Task [3] that was held in the framework of the Semeval-2007 Evaluation Workshop [2]. This task aims to disambiguate person names in Web search results. The participant's systems receive as input web pages retrieved from a Web search engine using an ambiguous person name as a query (e.g. "John Smith"). The systems output must specify how many referents (different people) exist for that person name and assign to each referent its corresponding documents. The challenge is to correctly estimate the number of referents and group documents referring to the same individual. A special characteristic of this test set is that many elements (web pages in the search results) are unique in their category. That is to say, many people are mentioned in just one web page. This means that a default strategy of locating each document in an individual cluster (maximum Purity) might also give an acceptable Inverse Purity. That is, most of classes of documents will be totally covered in one cluster.

Figure 1 shows the Purity and Inverse Purity values obtained for each system. The figure shows that there is an important trade-off between Purity and Inverse Purity. As a result, depending on the metric weighting in the F combining function, the system are ranked in a different way. Figure 2 shows the F values obtained by each system across different α values in F. This graph includes two baseline systems that consist of producing one cluster for each document (B_1) and grouping all documents in one cluster (B_{100}).

[2]http://nlp.cs.swarthmore.edu/semeval

$F_{0.5}$		$F_{0.2}$	
Ranked systems	F result	Ranked systems	F result
S_1	0,78	S_1	0,83
S_2	0,75	S_3	0,78
S_3	0,75	S_2	0,77
S_4	0,67	S_6	0,76
S_5	0,66	S_5	0,73
S_6	0,65	S_8	0,73
S_7	0,62	S_{11}	0,71
$\mathbf{B}_1$	0,61	S_7	0,67
S_8	0,61	S_{14}	0,66
S_9	0,58	S_{15}	0,66
S_{10}	0,58	S_{12}	0,65
S_{11}	0,57	S_9	0,64
S_{12}	0,53	S_{13}	0,63
S_{13}	0,49	S_4	0,62
S_{14}	0,49	S_{10}	0,6
S_{15}	0,48	$\mathbf{B}_{100}$	0,58
$\mathbf{B}_{100}$	0,4	S_{16}	0,56
S_{16}	0,4	$\mathbf{B}_1$	0,49

Table 1: Rankings for F0.5 and F0.2 using Purity and Inverse Purity

Note that B_1 is better than most systems according to α values bigger than 0.5. We could conclude that most of systems do not behave better than assigning one cluster to each document (baseline B_1). But the nature of the task suggests a different α value for the evaluation. Let us consider two alternative clustering distributions. In the first one, all the relevant documents are included in the same cluster which contains also some non relevant documents. In the other clustering distribution, there exists a cluster containing just relevant documents, but not all of them. The first distribution will obtain the maximum Inverse Purity, while the second distribution will obtain the maximum Purity. According to $F_{0.5}$ both distributions are equivalent in terms of quality. However, from our point of view, the first distribution is better, given that it is easier to discard some document from the relevant cluster than exploring all clusters looking for the rest of relevant documents. In conclusion, the parameter α should be fixed, for instance, at 0.2, giving more weight to Inverse Purity than to Purity...

According to $F_{0.2}$ a different system ranking is obtained (see Table 1). In this case, the baseline B_1 does not improve any system. That is, according to $F_{0.2}$ most systems represent a contribution with respect to the baseline approaches. Our conclusion is that the task interpretation is crucial and it can affect substantially the results. In this case $F_{0.2}$ seems to be more reasonable for this particular task. However, why using $\alpha = 0.2$ and not, for instance, 0.3?

A standard statistical significance test (such as the T-test or Wilcoxon) does not address this issue, because it is only applied to the outcome variable F and does not consider

	B_1	S_{14}	Statistical significance
$F_{0.5}$	0,61	0,49	0,022
$F_{0.2}$	0,52	0,66	0,015

Table 2: Statistical significance of improvements: $F_{0.5}$ vs. $F0.2$

Purity and Inverse Purity values. For instance, B_1 improves S_{14} with statistical significance (see Table 2) according to the Wilcoxon test on $F_{0.5}$ ($\alpha < 0.05$) but it is improved with $F_{0.2}$. In addition, we have identified 105 system pairs where one system improves the other with statistical significance according to the Wilcoxon test. From this set, in 89 cases (%84) there exists a statistically significant quality decrease according to one of the metrics.

We might think that it is enough to use the same α parameterization that is used in the competition for which our system is designed. However, the meaning of the α value can change across competitions depending on the data distribution. For instance, according to $F_{0.5}$ the one-in-one baseline approach improved the all-in-one baseline for WePS-1. However, the situation reverses in WePS-2: the all-in-one baseline seems substantially better.

In summary, we need a metric combining function which does not depend on any arbitrary weighting criterion. This measure should ensure that a system improvement is robust across metric combining criteria and it should also reflect the range of the improvement in order to select the best one.

3. PROPOSAL

3.1 Unanimous Improvements

The problem of combining evaluation metrics is closely related with the theory of conjoint measurement. In [1] it is described in detail the role of conjoint measurement theory in our problem. Rijsbergen [10] argued that it is not possible to determine empirically which metric combining function (over Precision and Recall) is the most adequate in the context of Information Retrieval evaluation. However, starting from the measurement theory principles, Rijsbergen described the set of properties that a metric combining function should satisfy. This set includes the *Independence* axiom (also called *Single Cancellation*), from which the *Monotonicity* property derives. The Monotonicity axiom implies that the quality of a system that surpasses or equals another one according to all partial metrics is necessarily equal or better than the second. In other words, it represents an improvement with no dependence on how the metrics were weighted.

We will refer to this quality relation as an *Unanimous Improvement*. Formally, being $Q_X(a)$ the quality of a according to a combining function of metrics in X:

$$Q_X(a) \geq_\forall Q_X(b) \text{ if and only if } x(a) \geq x(b) \forall x \in X$$

This relation has no dependence on how metrics are scaled, weighted or on their degree of correlation in the metric set. In other words, it implies an "empirical" improvement, but without information about the improvement range. From its definition and the antisymmetry property, the equality ($=_\forall$) and the strict relationship $>_\forall$ are derived. The unanimous equality implies that both systems obtain the same score for all metrics:

$$Q_X(a) =_\forall Q_X(b) \equiv (Q_X(a) \geq_\forall Q_X(b)) \wedge (Q_X(b) \geq_\forall Q_X(a))$$

The strict unanimous improvement implies that one system improves strictly the other for all metrics:

$$Q_X(a) >_\forall Q_X(b) \equiv (Q_X(a) \geq_x Q_X(b)) \wedge \neg(Q_X(a) =_x Q_X(b)) \equiv$$

$$(Q_X(a) \geq_\forall Q_X(b)) \wedge \neg(Q_X(a) \geq_\forall Q_X(b))$$

The non comparability $\|$ is also derived. It means that some metrics reward one system and some metrics reward the other. We refer to this cases as *metric biased improvements*.

$$Q_X(a) \|_\forall Q_X(b) \equiv \neg(Q_X(a) \geq_\forall Q_X(b)) \wedge \neg(Q_X(b) \geq_\forall Q_X(a))$$

The theoretical properties of the Unanimous Improvement are described in depth in [1]. The most important is that the Unanimous Improvement is the only relational structure that, while satisfying the Independence (Monotonicity) axiom, does not depend on metric weightings. In other words, we can claim that: *A system improvement according to a metric combining function does not depend in any way on metric weightings only if there is no quality decrease according to any individual metric.*

3.2 Unanimous Improvement Ratio

Given that the Unanimous Improvement is the only metric combining function that does not depend on metric weighting, our unique observable over each test case is a three-valued function (unanimous improvement, equality or biased improvement). However, we need a quantitative function in order to validate system improvements.

Having two systems a and b and the Unanimous Improvement relationship over test cases, we have samples for which a improves b ($Q_X(a) \geq_\forall Q_X(b)$), b improves a ($Q_X(b) \geq_\forall Q_X(a)$) and biased improvements ($Q_X(a) \|_\forall Q_X(b)$). We will refer to these sets as $T_{a \geq_\forall b}$, $T_{b \geq_\forall a}$ and $T_{a \|_\forall b}$ respectively. The total amount of samples will be referred as T. We want to define the quantitative measure Unanimous Improvement Ratio (UIR) according to three formal restrictions:

1. An increment of $T_{a \|_\forall b}$ samples implies a decrement in MIR. In the extreme case, if all samples are metric weighting biased ($T_{a \|_\forall b} = T$) then UIR=0.
2. If $T_{a \geq_\forall b} = T_{b \geq_\forall a}$ then UIR= 0.
3. Given a fixed $T_{a \|_\forall b}$, UIR is proportional to $T_{a \geq_\forall b}$ and inversely proportional to $T_{b \geq_\forall a}$.

Given these restrictions, we propose the following UIR definition:

$$\text{UIR}_{X,T}(a,b) = \frac{|T_{a \geq_\forall b}| - |T_{b \geq_\forall a}|}{|T|} =$$

$$\frac{|t \in T / Q_X(a) \geq_\forall Q_X(b)| - |t \in T / Q_X(b) \geq_\forall Q_X(a)|}{|T|}$$

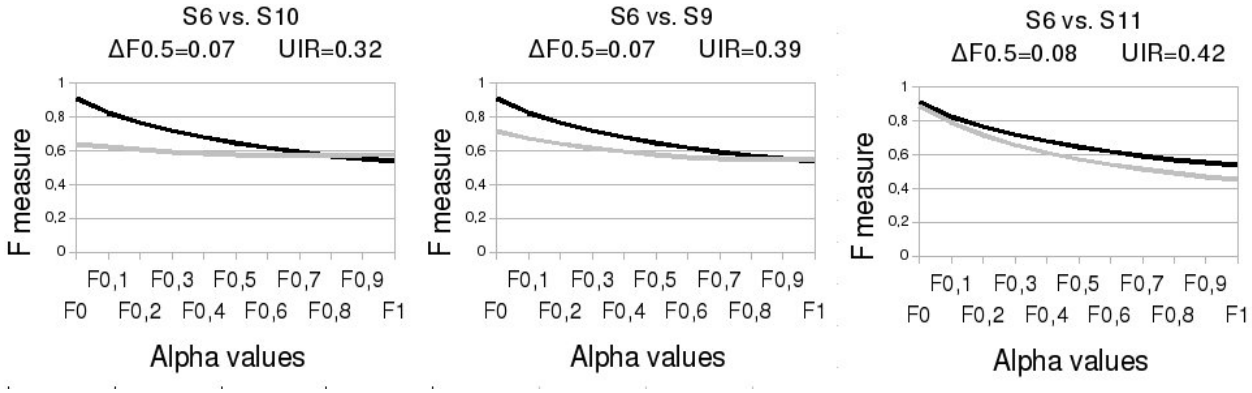

Figure 3: F measure vs. UIR: rewarding robustness

UIR has two main limitations. First, as well as the Unanimous Improvement, it is not *transitive* [1]. Therefore, it is not possible to define a linear system ranking based on UIR. In addition, there is some information loss when comparing systems given that the ranges in evaluation results are not considered.

On the other hand, the main advantage of UIR is that no metric weighting is necessary. In addition, given that the Unanimous Improvement does not consider metric ranges, the scale properties or normalization issues of individual metrics do not affect the results.

4. EMPIRICAL STUDIES

This section provides experiments in order to confirm that:

1. UIR rewards improvements that are robust across metric weighting schemes.
2. Given a set of equally robust improvements, the measure rewards the system that produces the largest improvement.
3. There exists a threshold for UIR values such that obtaining a UIR above the threshold guarantees that an improvement is robust, and this threshold is not too strong to identify differences between systems.

4.1 Rewarding Robustness across α values

Figure 3 shows three examples of system comparisons. Each curve represents the F_α value obtained for the system for different α values. System S6 (black curves) is compared with S10, S9 and S11 (grey curves) in each of the graphs. In all cases there is a similar quality increase according to $F_{0.5}$. However, UIR points out some differences: Depending on to what extent the improvement is robust across α values in F, UIR assigns diferent values to the improvement. S6 vs. S11 (rightmost graph) gives the largest UIR, because those systems do not swap their F values for any α. S6 vs. S10, on the other hand, has the smallest UIR value because the performances of S6 and S10 swap around $\alpha = 0.8$.

Another way of testing whether UIR rewards robustness is to consider separately two kinds of system comparisons: (i) system pairs for which F_α increases for all α values, and (ii) system pairs for which F increases for some α values and decreases for other α values. Table 4.1 shows the average increments for UIR and $F_{0.5}$ in each case. Note that UIR substantially rewards the absence of contradiction between α values (0.53 vs. 0.14). Notably, the absolute increase of $F_{0.5}$ is similar for both cases. In other words, although $F_{0.5}$ assigns the same relevance to purity and inverse purity, a certain $F_{0.5}$ improvement range does not say anything about whether we are being able to improve both purity and inverse purity at the same time.

	Improvements for all α 28 system pairs	Other cases 125 system pairs
$\lvert \triangle F_{0.5} \rvert$	0.12	0.13
|UIR|	0.53	0.14

Table 3: UIR and $F_{0.5}$ increase when F increases for all α values

	Robust improvements 53 pairs	Contradictory improvements 89 pairs	No imp. 11 pairs
$\lvert \triangle F_{0.5} \rvert$	0.11	0.15	0.05
|UIR|	0.42	0.08	0.027

Table 4: UIR and $F_{0.5}$ increases vs. statistical significance tests

We can also confirm this conclusion by considering independently both metrics (Purity and Inverse Purity). According to the statistical significance of the improvements for independent metrics, we can distinguish three cases:

1. *Contradictory improvements*: One metric increases and the other decreases, both with statistical significance.
2. *Robust improvements*: Both metrics improve significantly, or at least one improves significantly and the other does not decrease significantly.
3. *No improvement*: There is no statistically significant differences for any metric.

We use for this purpose the Wilcoxon test with $p < 0,05$. Surprisingly, Table 4.1 shows that the $F_{0.5}$ increase is even bigger when improvements are contradictory than when they are robust. Apparently $F_{0.5}$ rewards individual metric improvements obtained at the cost of (smaller) decreases in

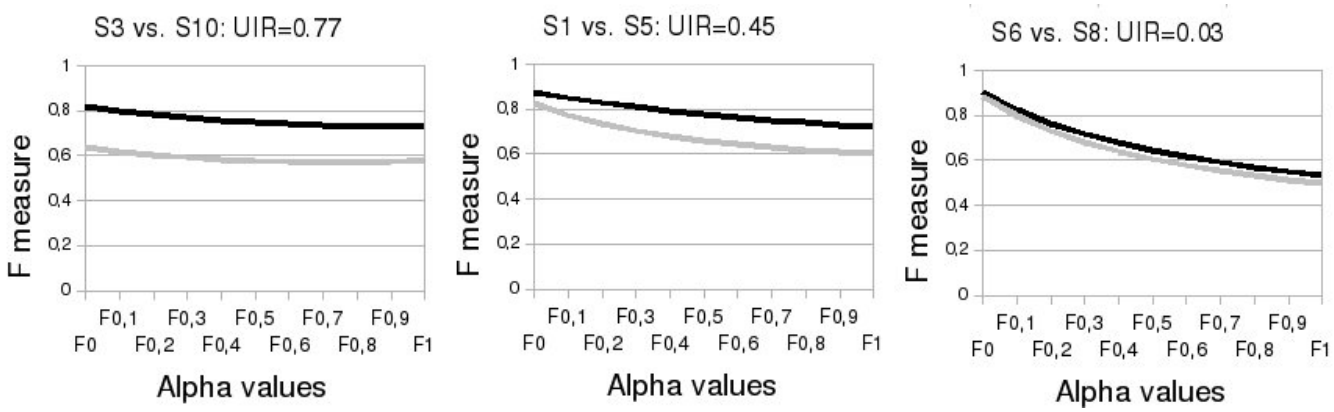

Figure 4: F vs. UIR: reflecting improvement ranges

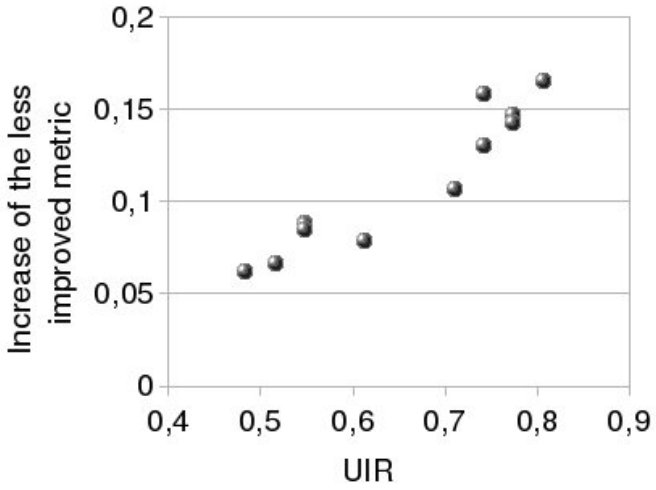

Figure 5: UIR vs. the improvement according to the less improved metric.

the other metric. UIR has a sharply different behaviour, rewarding robust improvements.

4.2 Reflecting Improvement Ranges

We have empirically verified that UIR reflects to what extent an improvement is robust across alternative α values. However, given a set of equally robust improvements, the measure should also reward the system that produces the largest improvement.

Let us consider an example taken from the WePS-1 testbed. Figure 4 represents the $F_{\alpha\in[0,1]}$ values for three system pairs. In all cases one system improves the other for all α values, but depending on the improvement range, UIR assigns higher values to larger improvements.

In fact, when both metrics are improved, the metric that has the weakest improvement determines the behaviour of UIR: Figure 5 illustrates this relationship for the ten system pairs with a largest improvement for both criteria; the Pearson correlation in this graph is 0.94.

4.3 UIR Threshold

What UIR value is appropriate to state that a system improvement is robust enough? We could set a very restrictive threshold and say, for instance, that an improvement is significantly robust when UIR$\geq$ 0.75. But such restriction would hardly be satisfied, and then the UIR test would not be informative: many robust system improvements would remain undetected by this test.

So our question now is whether there exists a threshold for UIR values such that obtaining a UIR above the threshold guarantees that an improvement is robust, and at the same time the threshold is not too strong to identify actual differences between systems.

Figure 6 shows the ratio of system pairs (a, b) (black curve) such that UIR(a, b) is bigger than a given threshold (horizontal axis). We have added a few more curves that represent key features of the system pairs:

- The proportion of robust system improvements,i.e. cases where both metrics improve significantly, or at least one improves significantly and the other does not decrease significantly
- The proportion of contradictory system improvements (see definition above).
- The ratio of system pairs for which $F_{0.5}$ increases for all α values ($F_\alpha(a) > F_\alpha(b) \forall \alpha$).

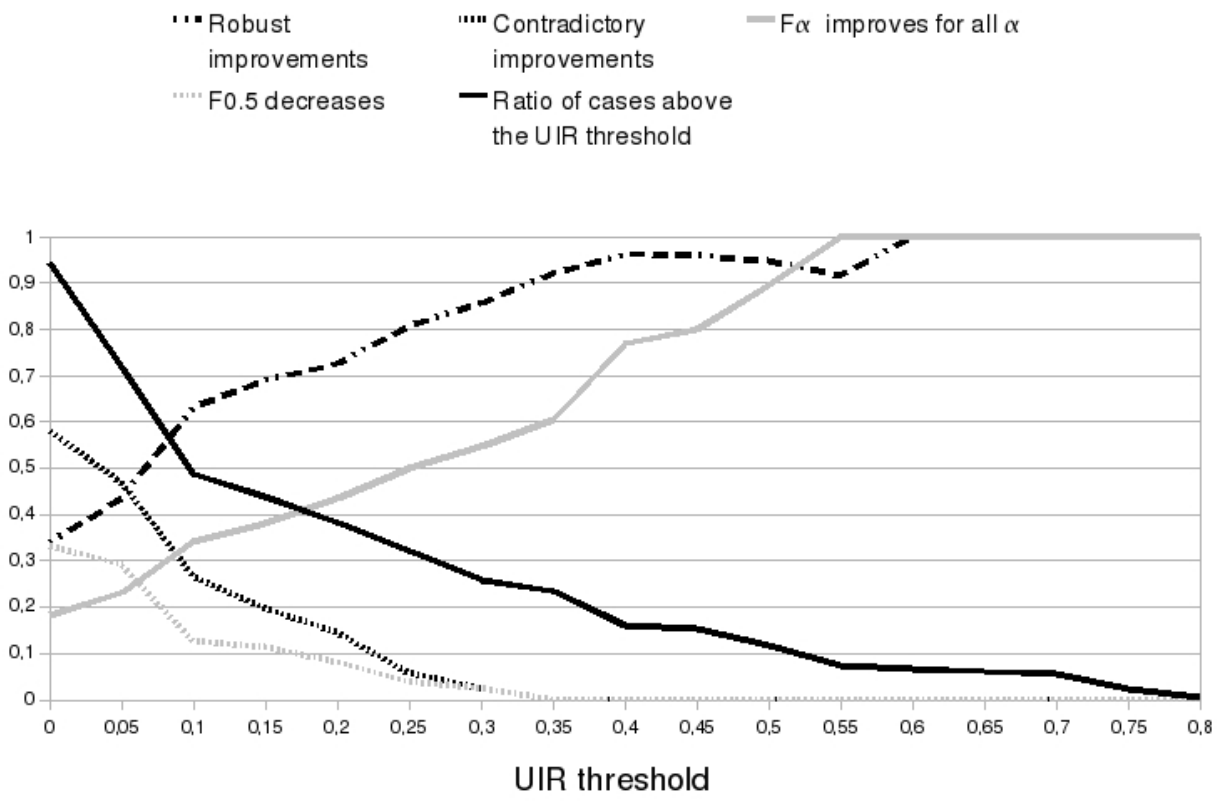

Figure 6: Improvement detected accross UIR thresholds

- The ratio of system pairs for which $F_{0.5}$ decreases although UIR is positive ($F_{0.5}(a) < F_{0.5}(b)$).

As the figure shows, an UIR threshold of 0.25 accepts around 25% of all system pairs. From this set, the number of contradictory improvements and the number of cases where $F_{0.5}$ decreases are low (%4 and %6 respectively). Also, in 50% of the cases F_α increases for all α values, and in 80% of the cases improvements are robust. It seems, therefore, that UIR$\geq$ 0.25 is a reasonable threshold.

5. USE CASE: THE WEPS-2 DATASET

In order to illustrate how UIR can be used, we analyze the results of the WePS-2 evaluation campaign [4], where BCubed Precision and Recall metrics were used. The official system ranking was generated according to $F_{0.5}$. The best run for each participant was included in the final ranking. In addition, three baseline approaches were included: all documents in one cluster (B_{100}), each document in one cluster (B_1) and the union of both (B_{COMB}) - see [4] for an explanation -.

Table 5 shows the results of applying UIR to the WePS-2 systems. The third column represents the set of systems that are improved by the corresponding system with a UIR>0.25. The fourth column represents the *reference system*, defined as, given a system a, the system that improves a with maximum UIR. It represents the system with which a should be replaced in order to improve results without sacrificing any partial evaluation metric. Finally, the last column represents the UIR between the system and its reference.

Note that UIR adds new insights into the evaluation process. First of all, note that, although the three top-scoring systems have a similar performance in terms of F (0,82, 0,81 and 0,81), PolyUHK is consistently the best according to UIR (it is the reference for 10 systems). In the most extreme case, UIR(PolyUHK,PRIYAVEN)=1, which means that PolyUHK improves both precision and recall of PRIYAVEN for all test cases in the dataset. Therefore, UIR clearly points out a best system, where F alone could only discern a set of three top scoring systems.

Note also that, although the ALL_IN_ONE baseline is better than five systems according to F, it is not better than any of them according to UIR. In fact, only the ONE_IN_ONE baseline is able to improve some system (BUAP_2). Therefore, UIR also adds the capability of detecting baseline approaches: if a system is adopting a baseline behaviour (for instance, using a very low clustering threshold that ends up setting up one big cluster), F will not clearly signal this problem (the F value obtained is better than five systems), but UIR will signal a problem, because this baseline strategy is not able to robustly improve any system.

6. CONCLUSIONS

The analysis described in this paper shows that the comparison of systems in clustering tasks is highly sensitive to the way of combining evaluation metrics. The UIR measure presented in this paper allows to combine evaluation metrics without assigning a relative weight to each metrics, and the empirical analysis has showed that UIR rewards robust improvements with respect to different metric weights.

System	$F_{0.5}$	Improved systems (UIR > 0.25)	Reference system	UIR for the reference system
PolyUHK (S1)	0,82	S2 S4 S6 S7 S8 S11..S17 B_1	-	-
ITC-UT_1 (S2)	0,81	S4 S6 S7 S8 S11..S17 B_1	S1	0,26
UVA_1 (S3)	0,81	S2 S4 S7 S8 S11..S17 B_1	-	-
XMEDIA_3 (S4)	0,72	S11 S13..S17	S1	0,58
UCI_2 (S5)	0,71	S12..S16	-	-
UMD_4 (S6)	0,71	S4 S7 S11 S13..S17 B_1	S1	0,35
FICO_3 (S7)	0,70	S11 S13..S17	S2	0,65
LANZHOU_1 (S8)	0,70	S11..S17	S1	0,74
UGUELPH_1 (S9)	0,63	S4 S12 S14 S16	-	-
CASIANED_5 (S10)	0,63	S12..S16	-	-
AUG_4 (S11)	0,57	S14..S17	S3	0,68
UPM-SINT_1 (S12)	0,53	S14 S16	S1	0,71
ALL_IN_ONE_BASELINE (B_{100})	0,53	B_{COMB}	-	-
UNN_2 (S13)	0,52	S15 S16	S1	0,9
COMBINED_BASELINE (B_{COMB})	0,52	-	B_{100}	0,65
ECNU_1 (S14)	0,42	-	S1	0,9
UNED_3 (S15)	0,41	S16	S1	0,97
PRIYAVEN (S16)	0,39	-	S1	1,00
ONE_IN_ONE_BASELINE (B_1)	0,34	S17	S1	0,29
BUAP_2 (S17)	0,33	-	S6	0,84

Table 5: WePS-2 results with Bcubed Precision and Recall, F and UIR measures.

UIR can be exploited in two ways. First, according to the UIR$\geq$ 0.25 threshold that was inferred from our empirical study, UIR is able to test the robustness of system improvements in shared tasks (such as the WePS clustering task). Second, given that UIR provides quantitative values, it is an alternative way of selecting the best approach during system training processes.

An UIR evaluation package is available for download at http://nlp.uned.es

Acknowledgments

This work has been partially supported by a grant from the Spanish government (project Text-Mess) and a grant from the European Commission (project TrebleCLEF).

7. REFERENCES

[1] E. Amigó, J. Gonzalo, and J. Artiles. Combining Evaluation Metrics via the Unanimous Improvement Ratio and its Application to Clustering Task. Technical Report, Departmento de Lenguajes y Sistemas Informáticos, Universidad nacional de Educacion a Distancia, Madrid, Spain. http://nlp.uned.es, 2009.

[2] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 2008.

[3] J. Artiles, J. Gonzalo, and S. Sekine. The Semeval-2007 Weps Evaluation: Establishing A Benchmark For The Web People Search Task. In *In Proceedings Of The 4th International Workshop On Semantic Evaluations (Semeval-2007)*, 2007.

[4] J. Artiles, J. Gonzalo, and S. Sekine. WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. In *In Proceedings Of The WePS-2 Workshop WWW-2009, Madrid 2009*, 2009.

[5] A. Bagga and B. Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85, 1998.

[6] J. Ghosh. Scalable clustering methods for data mining. In N. Ye, editor, *Handbook of Data Mining*. Lawrence Erlbaum, 2003.

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[8] M. Meila. Comparing clusterings. In *Proceedings of COLT 03*, 2003.

[9] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques, 2000.

[10] C. Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

[11] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. Technical Report TR 01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001.