Pattern Recognition Letters 167 (2023) 115–121

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Soft precision and recall

Pasi Fränti[a,*], Radu Mariescu-Istodor[b]

[a] *School of Computing, University of Eastern Finland, Joensuu 80101, Finland*
[b] *Karelia University of Applied Sciences, Joensuu 80200, Finland*

ARTICLE INFO

Article history:
Received 21 July 2022
Revised 14 December 2022
Accepted 2 February 2023
Available online 10 February 2023

Editor: Edited by Prof. S. Sarkar

Keywords:
Soft measures
Evaluation
Precision
Recall
F-score

ABSTRACT

Precision and recall are classical measures used in machine learning. However, they are based on exact matching. This results in binary classification where the predicted item is either a true or false positive despite inexact matching is often preferred in pattern recognition. To address this problem, we introduce soft variants of precision and recall based on application-specific similarity measure.

1. Introduction

Precision and *recall* are the most used measures to evaluate performance in various information retrieval and pattern recognition applications [1]. They provide scores in range [0,1] for a set of predicted items P in respect to a ground truth G.

Precision and recall are defined using set theory, see Fig. 1. *Precision* is the number of correct results (*true positives*) relative to the number of all results. *Recall* is the number of correct results relative to the number of expected results:

$$Precision = \frac{|G \cap P|}{|P|} \tag{1}$$

$$Recall = \frac{|G \cap P|}{|G|} \tag{2}$$

The performance is a trade-off between the precision and recall. Recall can be increased by lowering the selection threshold to provide more predictions at the cost of decreased precision. The performance can be unified into a single value called *F-score*:

$$F - score = 2\frac{|G \cap P|}{|G| + |P|} \tag{3}$$

The main assumption is that the predictions and the ground truth come from the same finite itemset. This set is often small, and the choices are mutually exclusive. However, this assumption does not hold in many applications. For instance, the precision in Fig. 2 is only 2/5 = 40% even if there is only one irrelevant result (*plaza*).

Another example is *speech activity detection* where the goal is to segment an audio file into *speech* and *non-speech* segments [2]. Sometimes multiple classes such as *speech*, *mouse click*, and *keyboard typing* are used [3]. Precision and recall based measures have been used but the problem is how to match the segments. For example, every detected event in Fig. 3 has its counter-part in the ground truth. Simple matching would therefore result in 1.0 precision and recall values despite the detection is far from perfect.

Event-based evaluation was considered by Measoros et al. [3] by allowing only a degree of misalignment between the ground truth and the predicted segments. However, one then needs to define what constitutes correct and incorrect detection. They defined that an event was labeled correctly if there was even a smallest temporal overlap with a detected event with the same label. This kind of difficulties have forced researchers to use-application specific measures, or to use ad hoc ways for measuring precision and recall [5]. Many researchers simply avoid the problem and use only frame-level measures with short 20 ms frames [2].

To address the above problems, we introduce *soft precision* and *soft recall* to the above-mentioned applications and beyond. The measures are based on the *soft cardinality* by Jimenez [7] and Jimenez et al. [6]. We demonstrate the proposed measures in two applications: keyword extraction [8], and retrieval of GPS trajectories [9]. In keyword extraction we use both syntactic [10] and semantic [11,12] text similarity measures.

[*] Corresponding author.
E-mail address: pasi.franti@uef.fi (P. Fränti).

https://doi.org/10.1016/j.patrec.2023.02.005

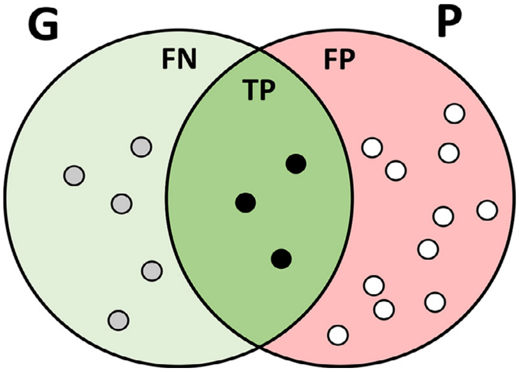


Fig. 1. Ground truth (G) and the predicted result (P). Precision and recall are defined based on true positives (TP), false positives and false negatives (FN). True negatives (TN) are usually ignored.

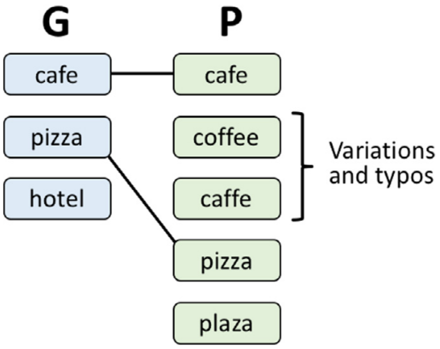


Fig. 2. Prediction (P) includes two correct items (*café* and *pizza*), one variation (*coffee*), and one with typing error (*caffe*).

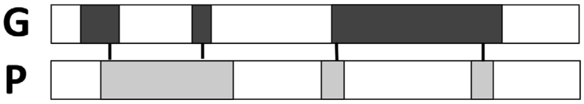


Fig. 3. Matching events based on segment overlap.

2. Approaches using inexact matching

Existing solutions have attacked the problem by relaxing the requirement for what is considered as a correct result. Ziolko [4] defined *fuzzy precision* and *recall* for audio signal segmentation by mapping the ground truth segments to the predicted result, and then summing up the similarities between the mapped segments.

A method called *similarity matching* [13] performs greedy pairing of the most similar pairs of items until one of the sets becomes empty. Remaining items are then matched with their most similar item allowing many-to-one mappings. Sum of the pairwise similarities are calculated as the overall score.

The method by Ziolko [4] performs one-directional mapping from G to P, whereas the method by Rezaei et al. [13] performs pairing, which provides a symmetric measure regardless which one is the ground truth. This provides only a single similarity score while Ziolko provides three values: precision, recall and F-score. In general, both address the problem as a set-matching problem, which closely resembles the cluster validation problem [14].

A completely different approach called *Active estimation of F-score* was proposed by Sawade et al. [15]. They consider the problem as lack of training data and aims at solving it by drawing new test instances.

In this paper, we follow the approaches by Ziolko [4] and Rezaei et al. [13] and utilize the similarities between the two item sets. In other words, we generalize the classical hard decision of precision and recall having binary value in {0,1} to a similarity measure having real value in range [0,1]. However, instead of calculating average similarity between the sets, we derive the two measures (precision and recall) from the pairwise similarities using a set theoretic definition.

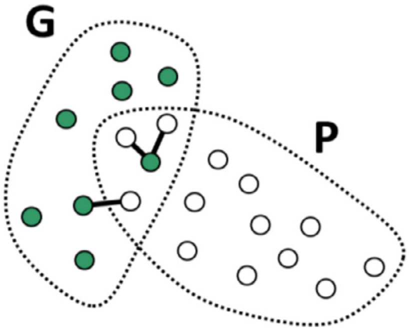


Fig. 4. Ground truth (G) and the prediction (P) in an imaginary feature space. True positives is a soft count of the soft intersection.

The proposed approach is demonstrated in Fig. 4 where none of the predicted items (white) is an exact match with any of the ground truth items (green). However, some are close and should contribute to the true positive count despite not being exact matches. To facilitate this idea, application specific similarity measure is applied. This changes the concept of how the true positives are calculated. G and P are no longer subsets of the same item set.

3. Soft precision and recall

Precision and recall measure the number of correctly predicted items relative to the number of all items predicted (*precision*), or relative to the number of all relevant items (*recall*).

The measures can be defined using the counts:

TP	True positives
FP	False positives
FN	False negatives

Alternative definition is based on set theory. Two different types of sets can be used: crisp and fuzzy. The classical set theory, denote here as *crisp sets*, uses binary membership values. *Fuzzy sets* define that an item belongs to a set with a certain degree, which is called a *membership value* in a range [0, 1].

However, neither of these definitions fits to our purpose. We have crisp sets since each item belongs only to one set. However, our items can be inexact and redundant. For example, the set {*cafe, café, cafeteria, coffee house, coffee shop*} includes five unique words but having high redundant content.

For the precision and recall, we do not need to know to what degree an item belongs to the set. Instead, we need to define their *counts*. For crisp sets the counts are trivially the number of items in the set, but not in our case. For instance, what should be the count for the *café* example in Fig. 2 with three truly unique keywords (*café, pizza, hotel*) with additional variations of the word coffee? We refer this as *cardinality* of the set. In crisp sets, cardinality is the number of *unique* items in the set.

3.1. Soft cardinality

In the segmentation example (Fig. 3), the cardinality of the set can be simply the length of the segment. In case of keywords (Fig. 2), however, different definition is needed. We adopt the *soft cardinality* by Jimenez et al. [6] based on a similarity relationship between the items. We first define *soft count* of an item as the inverse of the sum of its similarity to all other items in the set:

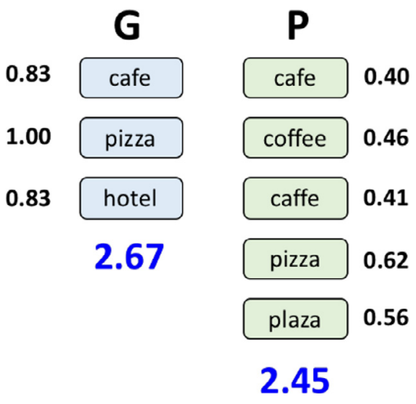


Fig. 5. Soft cardinalities of the Cafe example using the standard Levenshtein edit distance. For example, distances of *café* to the predicted words are: *café* (0), *coffee* (2), *caffe* (1), *pizza* (5), *plaza* (4).

$$count(A_i) = \frac{1}{\sum_{j=1}^{K} Sim\left(A_i,\ A_j\right)} \tag{4}$$

where A_i is an item in the set, and K is the size of the set. If an item is unique, it is similar only to itself and the count equals to 1.00. If there are duplicate items, their corresponding counts equals to 0.50. All similar items reduce the count accordingly. The cardinality of the set is the sum of the counts of the individual items:

$$card(A) = \sum_{i=1}^{K} count(A_i) \tag{5}$$

Distance measures can also be used by converting them first to a similarity measure. In case of Levenshtein edit distance, we use the conversion:

$$Sim\ \left(A_i, A_j\right) = 1 - \frac{edit\left(A_i,\ A_j\right)}{\max\left(|A_i|,\ \left|A_j\right|\right)} \tag{6}$$

An example using Levenshtein similarity is in Fig. 5. Here the crisp cardinalities are 3 (G) and 5 (P). However, the words *cafe, coffee* and *caffe* contribute much less than their hard count would indicate (3). Their soft count sums up only to 1.27, and the corresponding soft cardinality is therefore only 2.45; compared to the crisp cardinality of 5.00.

3.2. Union and intersection

For calculating precision and recall, we also need the cardinality of intersection. In the segmentation example, it is the length of the two intersecting segments. In other cases, we get intersection indirectly using union of the two sets. In crisp variant, it is the sum of the counts of the two sets.

In soft variant, we derive soft cardinality by taking the concatenation of the sets [6]. Doing this may create duplicates, however, the soft cardinality takes care of these naturally by halving their respective weights (soft counts).

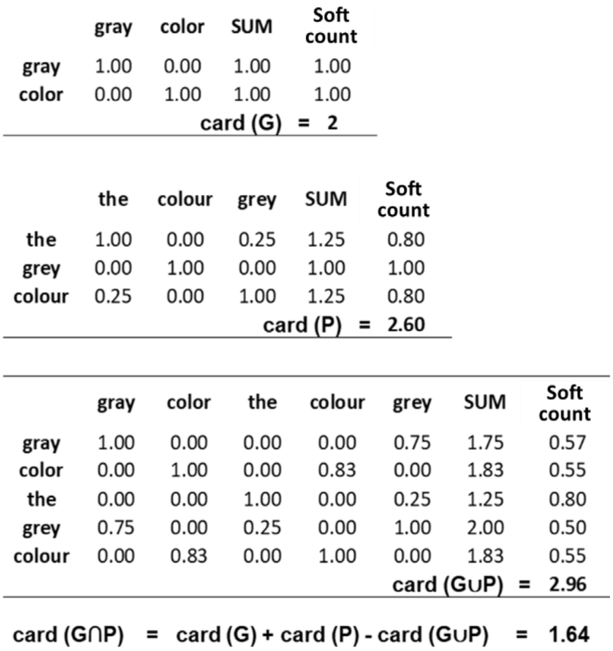
The intersection of two sets using soft cardinality can now be defined via the union as follows:

$$card(G \cap P) = card(G) + card(P) - card(G \cup P) \tag{7}$$

Fig. 6 shows an example of the calculation.

3.3. Precision and recall

Soft variants of the precision and recall can now be defined using the soft cardinalities of the sets and their intersection as shown in Table 1. The main difference to the crisp counterparts is the way the cardinalities are defined. Note that soft variants cannot anymore be defined based on the counts (TP, FP and FN).



	gray	color	SUM	Soft count
gray	1.00	0.00	1.00	1.00
color	0.00	1.00	1.00	1.00
		card (G)	=	2

	the	colour	grey	SUM	Soft count
the	1.00	0.00	0.25	1.25	0.80
grey	0.00	1.00	0.00	1.00	1.00
colour	0.25	0.00	1.00	1.25	0.80
			card (P)	=	2.60

	gray	color	the	colour	grey	SUM	Soft count
gray	1.00	0.00	0.00	0.00	0.75	1.75	0.57
color	0.00	1.00	0.00	0.83	0.00	1.83	0.55
the	0.00	0.00	1.00	0.00	0.25	1.25	0.80
grey	0.75	0.00	0.25	0.00	1.00	2.00	0.50
colour	0.00	0.83	0.00	1.00	0.00	1.83	0.55
					card (G∪P)	=	2.96

card (G∩P) = card (G) + card (P) - card (G∪P) = 1.64

Fig. 6. Example of calculating soft cardinality using Word2Vec semantic similarity.

Table 1
Crisp and soft variants with two notations.

Measure	Sets	Counts
Crisp		
Precision	$\frac{\lvert G\cap P\rvert}{\lvert P\rvert}$	$\frac{TP}{TP+FP}$
Recall	$\frac{\lvert G\cap P\rvert}{\lvert G\rvert}$	$\frac{TP}{TP+FN}$
F-score	$2\,\frac{\lvert G\cap P\rvert}{\lvert G\rvert+\lvert P\rvert}$	$2\,\frac{Precision\cdot Recall}{Precision+Recall}$
Soft		
Precision	$\frac{card(G\cap P)}{card(P)}$	N/A
Recall	$\frac{card(G\cap P)}{card(G)}$	N/A
F-score	$2\,\frac{card(G\cap P)}{card(G)+card(P)}$	N/A

Table 2
Description of the test examples.

Test set	Type	Language
Café	Syntactic	English US
University	Semantic	English US
Gray Color	Semantic	English US + UK

3.4. Semantic similarity

So far, we have considered only a syntactic measure, namely edit distance. In some cases, the semantic meaning of the words matters more than their typographic differences. We therefore consider also *Word2Vec* [11,12] for semantic similarity. Fig. 7 shows an example of the calculation of soft cardinality and corresponding measures.

4. Numerical examples

We first compare the soft and crisp measures using the setup shown in Table 2. Levenshtein edit distance is used for syntactic similarity and *Word2Vec* for semantic similarity trained for English

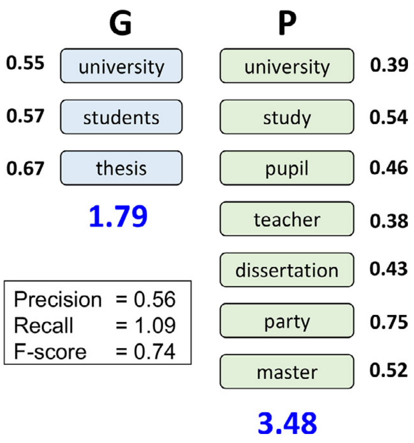


Fig. 7. Example of calculating soft cardinalities.

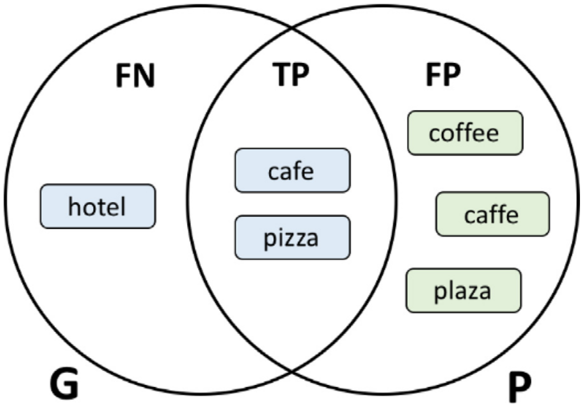


Fig. 8. Summary of the Cafe example.

Table 3
Cardinalities of set P and G in examples.

Set	Crisp	Syntactic	Semantic
Café			
G	3.00	2.67	1.83
P	5.00	2.45	2.14
G∪*P*	6.00	2.91	2.27
G∩*P*	2.00	2.21	1.70
University			
G	3.00	1.90	1.79
P	7.00	3.69	3.48
G∪*P*	9.00	3.88	3.32
G∩*P*	1.00	1.71	1.95
Gray Color			
G	2.00	2.00	1.39
P	3.00	2.60	3.00
G∪*P*	5.00	2.96	4.04
G∩*P*	0.00	1.64	0.35

language by Google.[1] Web interface of the soft measures is publicly available.[2] It supports several syntactic measures and the semantic similarity measure. Summary of the examples is presented in Figs. 8–10 and the cardinalities of the sets are summarized in Table 3.

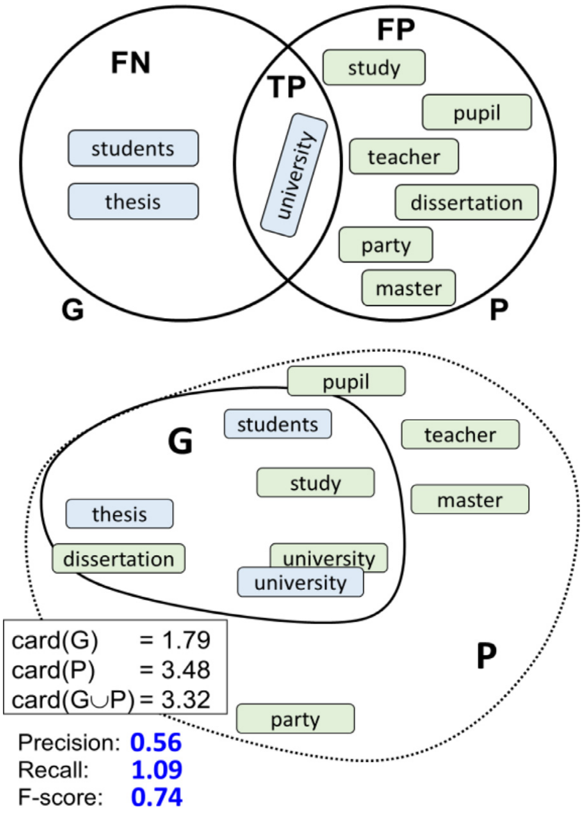


Fig. 9. Summary of the University example.

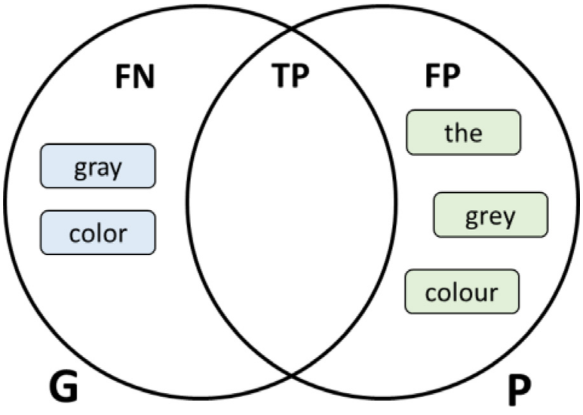


Fig. 10. Summary of the Gray Color example.

4.1. Cafe example

Cafe example contains words from a business that offers accommodation, food, and cafeteria services. The predicted keywords catch two out of the three ground truth words, so the crisp recall (0.67) is correct. Semantic measure over-estimates the recall (0.93) because *plaza* is partly predicted by the other words. Besides this, all measures give reasonable results.

The precision values, however, show the need for the soft measure. Two of the predicted keywords have only slight syntactic deviations. *Coffee* and *caffe* (*café* miss-spelled) are close to the ground truth but only the soft variants provide high precision (0.80 and 0.90).

[1] https://code.google.com/archive/p/word2vec.

[2] https://cs.uef.fi/sipu/soft/SoftEval/.

4.2. University example

The crisp measures consider only *university* as a correct prediction resulting to a low F-score (0.14). Soft measures provide better scores (0.61 and 0.74). Especially the recall should be high as all the ground truth words are predicted in one form or another. Semantic variant works better because it can recognize the similarities between the words *university, teacher, student*, and *pupil*.

4.3. Gray color example

The third example has the correct prediction with only minor differences. Crisp variant fails with 0.00 scores for precision and recall. Surprisingly, the semantic variant also gives low scores. The reason originates to the *Word2vec* model which was trained using US English and cannot provide values for *color* and *gray* (British spellings).

The syntactic measure gives high recall (0.82) but cannot omit the stop word (*the*), which would require using NLP. As a result, the precision remains a bit lower than expected (0.63) but it is still the only measure that provides non-zero score.

4.4. Countries dataset

We next perform more extensive experiment using D6 from the *Countries dataset*.[3] It contains clusters of ten European countries: *Cyprus, France, Greece, Latvia, Monaco, Norway, Poland, Russia, Serbia* and *Sweden*. Each cluster contains 50 variations of the corresponding country name made by random character modifications (add, remove, substitute). For example, Cyprus clusters contain words like *oypmreus, cyprqs, cypcrus, yprus*.

In the experiment, we select three countries as the ground truth and make random predictions as follows. A correct prediction is to pick a random name from the cluster of the same country name. Four types of predictions are made: (1) pick all three correct; (2) pick only two correct; (3) pick all three correct and one incorrect; (4) pick two correct and one incorrect. Examples with the expected precision and recall values are (*Datia* is from the *Latvia* cluster):

GT:	Cyprys, France, Greece	**Prec.**	**Rec.**
1:	Oypmreus, Drance, Rerece	1.00	1.00
2:	Oypmreus, Drance	1.00	0.67
3:	Oypmreus, Drance, Rerece, **Datia**	0.75	1.00
4:	Oypmreus, Rerece, **Datia**	0.67	0.67

The experiments were repeated 100 times. We consider the soft variant with Levenshtein distance. Since the crisp variant is basically useless (all results 0), we also consider a simple thresholding variant (Crisp*) which match two strings if their Levenshtein similarity is 0.8 or higher. Average results in Table 4 show that the soft variant is clearly superior to the crisp variants.

4.5. Segmentation example

The next example is motivated by the speech segmentation. Here the similarity is measured simply by the length of the shared segments. Fig. 12 shows ground truth segments (blue) and the predicted result (red). There are no exact matches, so the crisp variant is again useless. Crisp* does not work either as none of the predicted segments have similarity above 0.8 with the ground truth. However, soft precision and recall values are meaningful and straightforward to calculate.

Table 4
Summary of the countries experiment.

Precision				
Method	1	2	3	4
Crisp	0	0	0	0
Crisp*	0.48	0.32	0.48	0.32
Soft	0.87	0.89	0.71	0.69
Expected	1.00	1.00	0.75	0.67
Recall				
Method	**1**	**2**	**3**	**4**
Crisp	0	0	0	0
Crisp*	0.48	0.48	0.36	0.32
Soft	0.89	0.66	0.92	0.72
Expected	1.00	0.67	1.00	0.67

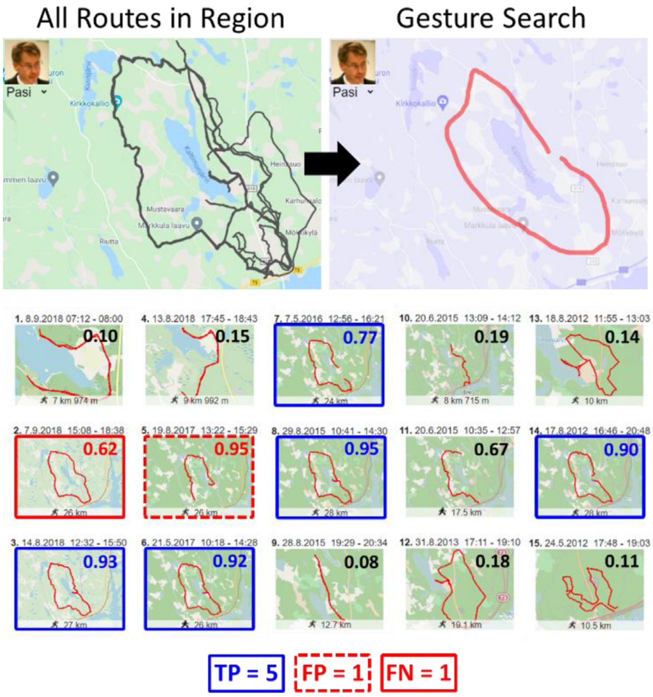


Fig. 11. Gesture search returns the most similar trajectories (C-SIM scores of 0.77 or higher in this case). The ground truth consists of the (almost) full hiking routes around lake *Kaltimo*, called *Kaltimonkierto*.

4.6. GPS trajectories

The final experiment is similarity-based search in Mopsi[4] where user provides a sample query shape drawn on map. The most similar trajectories are retrieved using a grid-based similarity measures [9] as shown in Fig. 11. In this case, we expect to receive full tours around the lake roughly following the so-called *Kaltimonkierto* route. The search is good but misses one tour following different path on the east side (0.62) and provides one partial tour (0.95).

The expected precision and recall values (0.83) are shown in Table 5 and Table 6 with the corresponding scores by the crisp and the proposed soft variant. The crisp variant is again useless (values 0) whereas crisp* provides correct precision of 0.83. Soft variant provides 0.86 and 0.95.

[3] https://cs.uef.fi/sipu/string/countries.

[4] https://cs.uef.fi/mopsi.

Expected	1.00	0.67	1.00	0.67

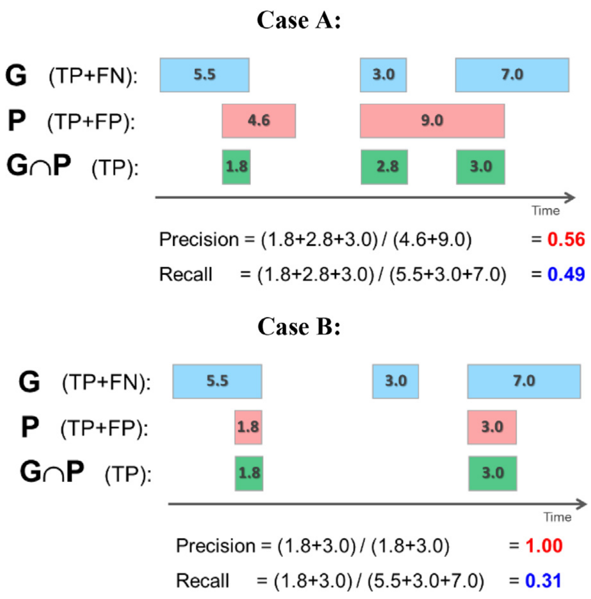

Case A:
G (TP+FN):
5.5
3.0
7.0
P (TP+FP):
4.6
9.0
G∩P (TP):
1.8
2.8
3.0
Time
Precision = (1.8+2.8+3.0) / (4.6+9.0) = 0.56
Recall = (1.8+2.8+3.0) / (5.5+3.0+7.0) = 0.49
Case B:
G (TP+FN):
5.5
3.0
7.0
P (TP+FP):
1.8
3.0
G∩P (TP):
1.8
3.0
Time
Precision = (1.8+3.0) / (1.8+3.0) = 1.00
Recall = (1.8+3.0) / (5.5+3.0+7.0) = 0.31

Fig. 12. Segmentation example.

Table 5
Summary of the segmentation experiment.

Case A Example	Precision	Recall	F-score
Crisp	0	0	–
Crisp*	0	0	–
FuzzyPR [4]	0.46	0.40	0.43
Soft	0.56	0.49	0.52
Case B **Example**	**Precision**	**Recall**	**F-score**
Crisp	0	0	–
Crisp*	0	0	–
FuzzyPR [4]	0.80	0.25	0.55
Soft	1.00	0.31	0.47

Table 6
Summary of the GPS trajectories experiment.

Example	Precision	Recall	F-score
Crisp	0	0	N/A
Crisp*	0.83	0.67	0.75
Soft	0.86	0.95	0.95
Expected	0.83	0.83	0.83

5. Conclusions

Soft variants for precision and recall were introduced based on application specific similarity measure. In case of strings, we considered both syntactic (Levenshtein) and semantic (Word2Vec) similarities. In case of GPS trajectories, we used grid-based similarity measure.

Experiments showed that the soft variants provide more meaningful result than the crisp variant. This can help researchers by reducing the time spent in creating ground truth datasets because the exactness requirement is no longer critical. The drawback of the method is that it requires application dependent similarity measure although natural choices exist at least in case of strings, segments, and GPS data.

The measure can also provide values greater that 1 in some cases. This originates from the definition of the soft cardinality, which itself is intuitively well defined but tends to over-estimate the set sizes. We considered alternative definitions but did not find one that would have worked perfectly in all situations. The cardinalities can nevertheless be always upper bounded if wanted.

Despite its deficiencies, the soft measures have shown its usefulness. It provides meaningful estimates for the keyword extraction [16] where the traditional crisp measures fail. It can be adopted to other information retrieval and pattern matching applications where inexact matches are used.

Authorship confirmation

1 This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2 If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3 If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4 All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5 All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript Eqs. (1)–((7)).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

References

[1] F. Pagani, M. Dell'Amico, D. Balzarotti, Beyond precision and recall: understanding uses (and misuses) of similarity hashes in binary analysis, in: Proceedings of the ACM Conf. on Data and Application Security and Privacy, 2018.
[2] A. Sholokhov, Md. Sahidullah, T. Kinnunen, Semi-supervised speech activity detection with an application to automatic speaker verification, Comp. Speech Lang. 47 (2018) 132–156.
[3] A. Mesaros, T. Heittola, T. Virtanen, Metrics for polyphonic sound event detection, Appl. Sci. 6 (2016) 162.
[4] B. Ziółko, Fuzzy precision and recall measures for audio signals segmentation, Fuzzy Sets Syst. 279 (2015) 101–111.
[5] Z. Wu, C. Su, M. Yin, Z. Ren, S. Xie, Subspace clustering via stacked independent subspace analysis networks with sparse prior information, Pattern Recognit. Lett. 146 (2021) 165–171.
[6] S. Jimenez, F. Gonzalez, A. Gelbukh, Text comparison using soft cardinality, in: Proceedings of the International Symposium on String Processing and Information Retrieval, 2010.
[7] S. Jimenez (2008): A knowledge-based information extraction prototype for data-rich documents in the information technology domain. Master's thesis, National University of Colombia.
[8] M. Rezaei, N. Gali, P. Fränti, Clrank: a method for keyword extraction from web pages using clustering and distribution of nouns, in: Proceedings fo the IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015.
[9] R. Mariescu-Istodor, P. Fränti, Gesture input for GPS route search, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, Cham, 2016, pp. 439–449.
[10] N. Gali, R. Mariescu-Istodor, P. Fränti, Framework for syntactic string similarity measures, Expert Syst. Appl. 129 (2016) 169–185.
[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[12] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems.

[13] M. Rezaei, P. Fränti, Matching similarity for keyword-based clustering. Structural, syntactic, and statistical pattern recognition. S+SSPR 2014, Lecture Notes in Computer Science, 8621, Springer, Berlin, Heidelberg, 2014.

[14] M. Rezaei, P. Fränti, Set matching measures for external cluster validity, IEEE Trans. Knowl. Data Eng. 28 (8) (2016) 2173–2186.

[15] C. Sawade, N. Landwehr, T. Scheffer, Active estimation of F-scores, in: Proceedings of the Int. Conf. on Neural Information Processing Systems (NIPS), 2010.

[16] H. Shah, P. Fränti, Combining statistical, structural, and linguistic features for keyword extraction from web pages, Appl. Comput. Intell. 2 (2) (2022) 115–132.