Proximity estimation and hardness of short-text corpora

Marcelo Luis Errecalde[1] Diego Ingaramo[1]
Paolo Rosso[2]

[1]Universidad Nacional de San Luis, Argentina
[2]Universidad Politécnica de Valencia, España

5th Int. Workshop on Text-based Information Retrieval, 2008

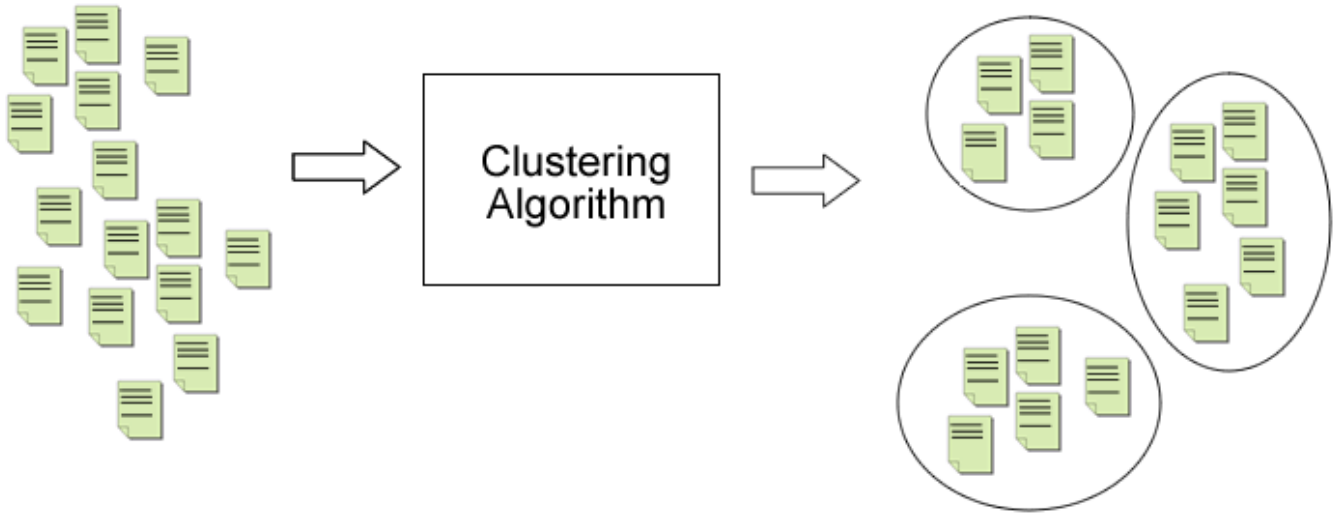
Outline

What is the problem we are working on?

1. Main goal: to develop effective algorithms for the problem of clustering short-text corpora.

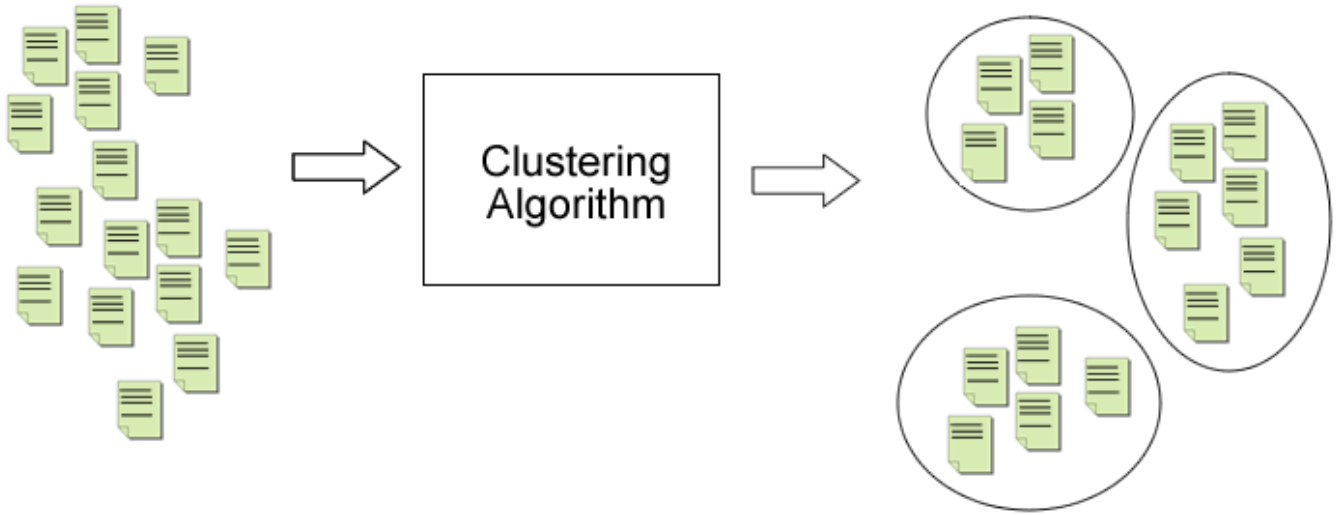
What is the problem we are working on?

1. Main goal: to develop effective algorithms for the problem of clustering short-text corpora.
2. These algorithms assign documents to unknown categories in an unsupervised way.



What is the problem we are working on?

1. Main goal: to develop effective algorithms for the problem of clustering short-text corpora.
2. These algorithms assign documents to unknown categories in an unsupervised way.



3. Our interest is on clustering of:
   - short-texts (in general)
   - narrow domain short-texts (in particular)

Why is it important?

- Applicability in different areas of text processing:
  - text mining
  - summarization
  - information retrieval
  - ...
- Tendencies of people to use 'small-languages':
  - blogs
  - text-messages
  - snippets
  - ...

Why is this problem difficult?

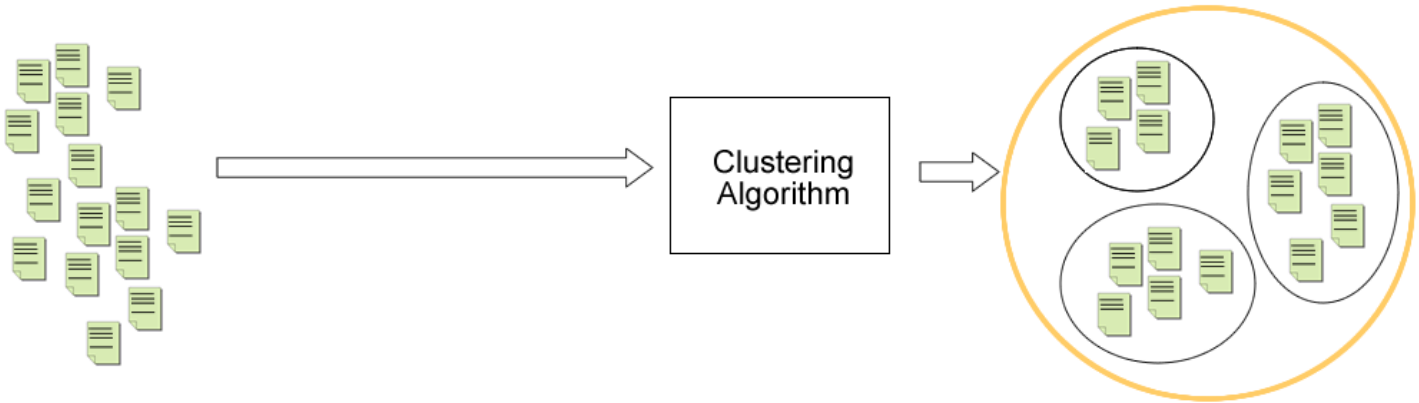
1. General problems of text clustering:
   - Synonymy.
   - Polysemy.
2. Additional difficulties due to:
   - Low frequencies of the document terms.
   - High overlapping degree of their vocabularies.

These aspects can negatively affect the estimation of how similar the documents are and (in consequence) the whole clustering process

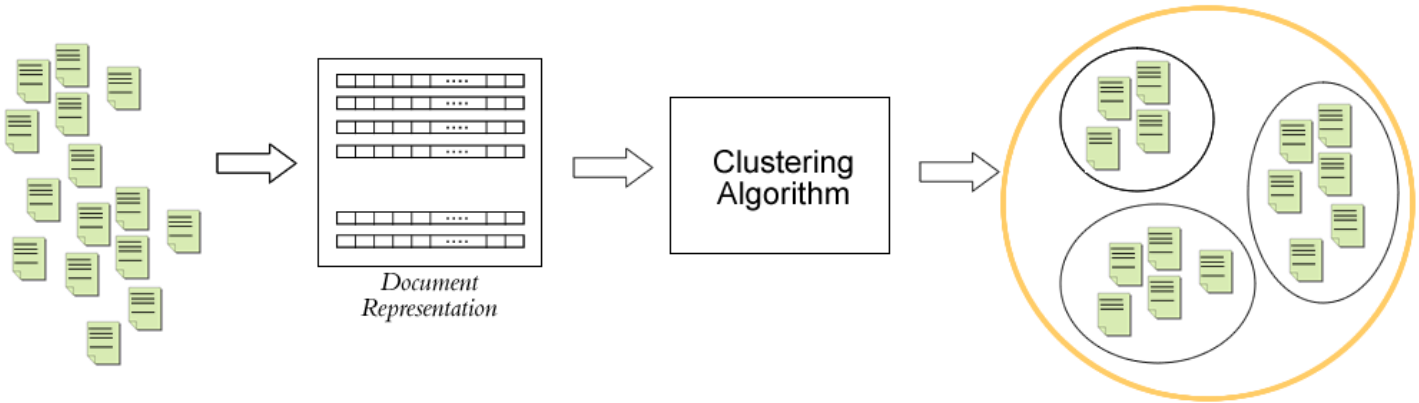
What questions are we trying to answer in our work?

1. it is usually assumed that short text corpora are harder to deal with than traditional corpora, but how harder?
2. how accurate traditional similarity measures in these cases are?
3. to what extent are both issues related?

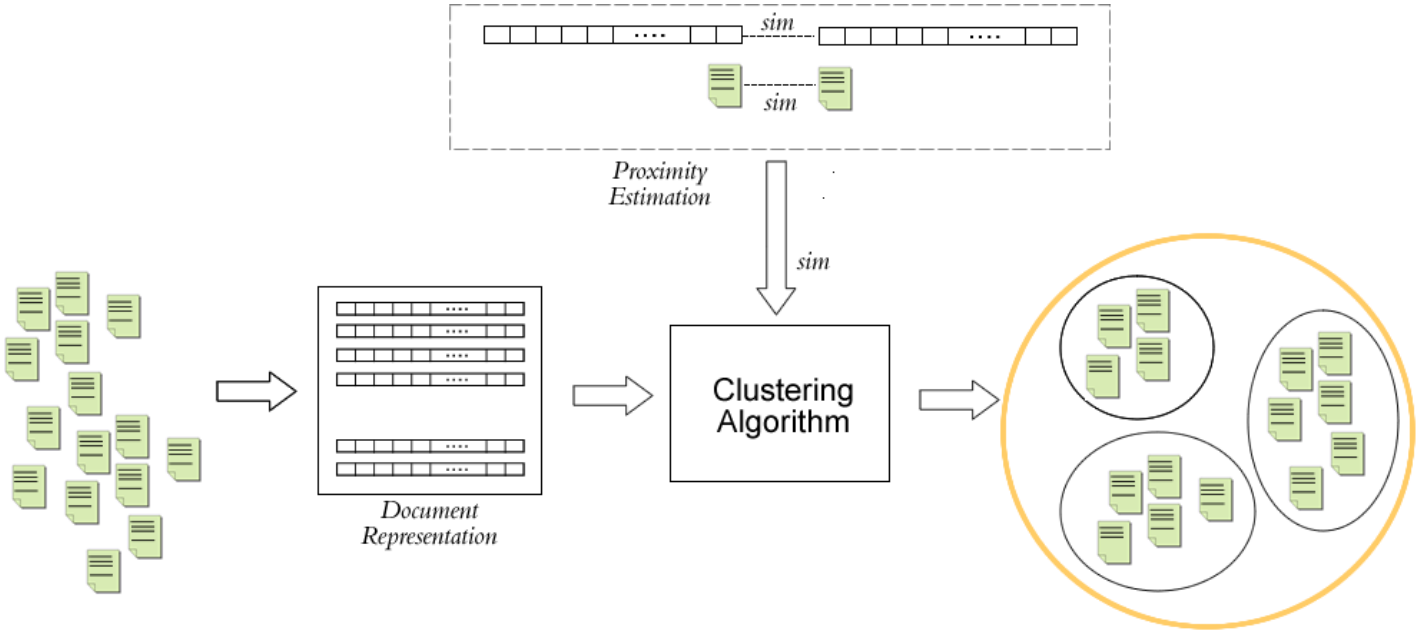
A more detailed look to the document clustering process...



A more detailed look to the document clustering process...

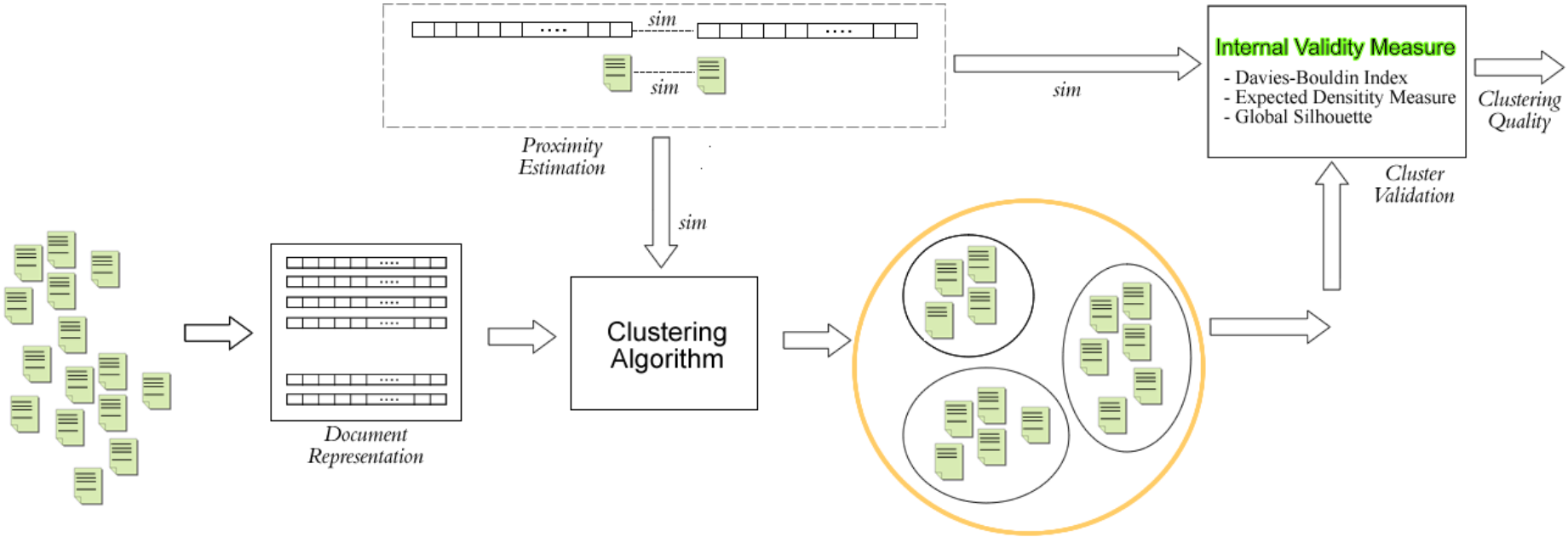


A more detailed look to the document clustering process...

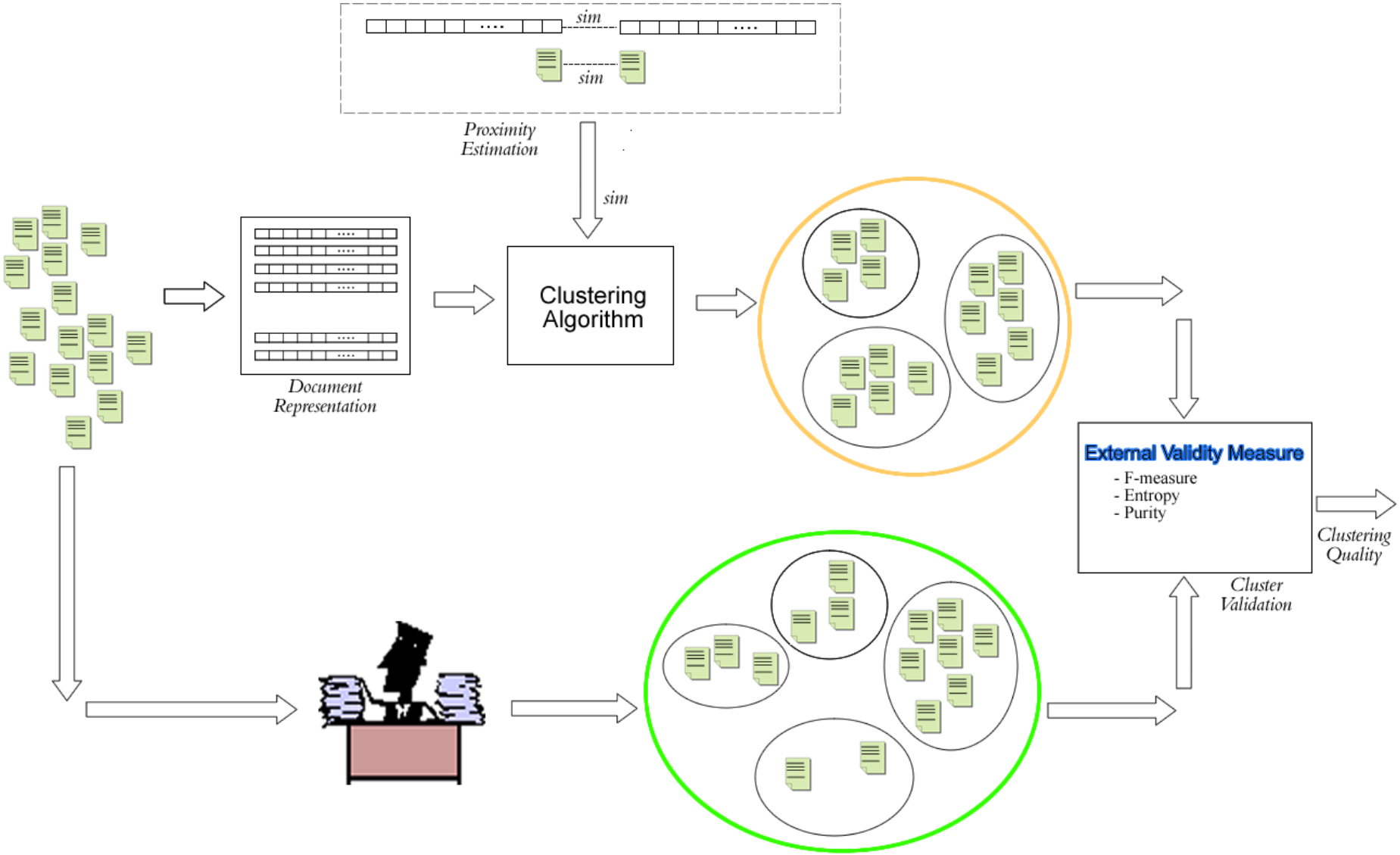


Clustering validation

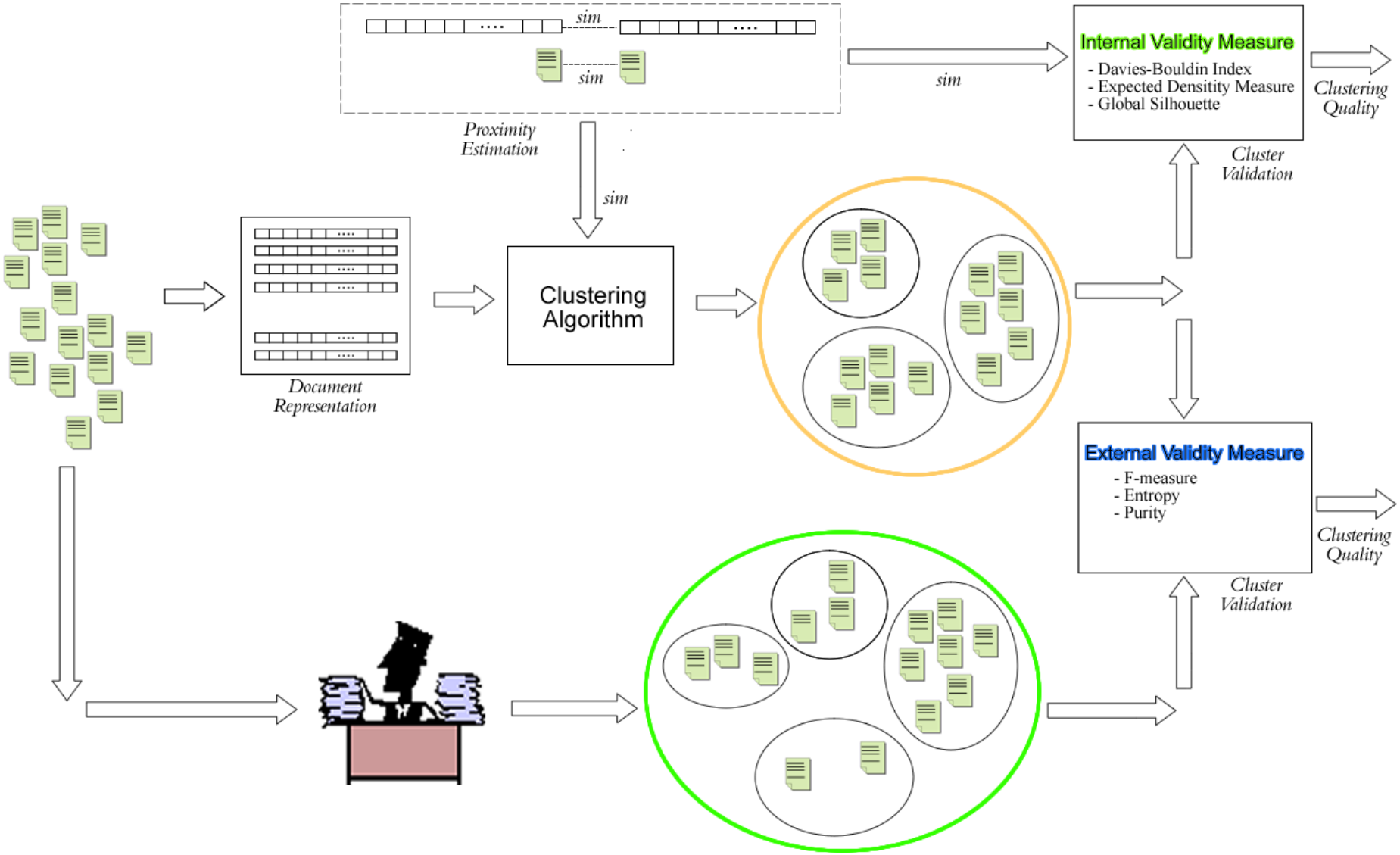
A more detailed look to the document clustering process...



A more detailed look to the document clustering process...

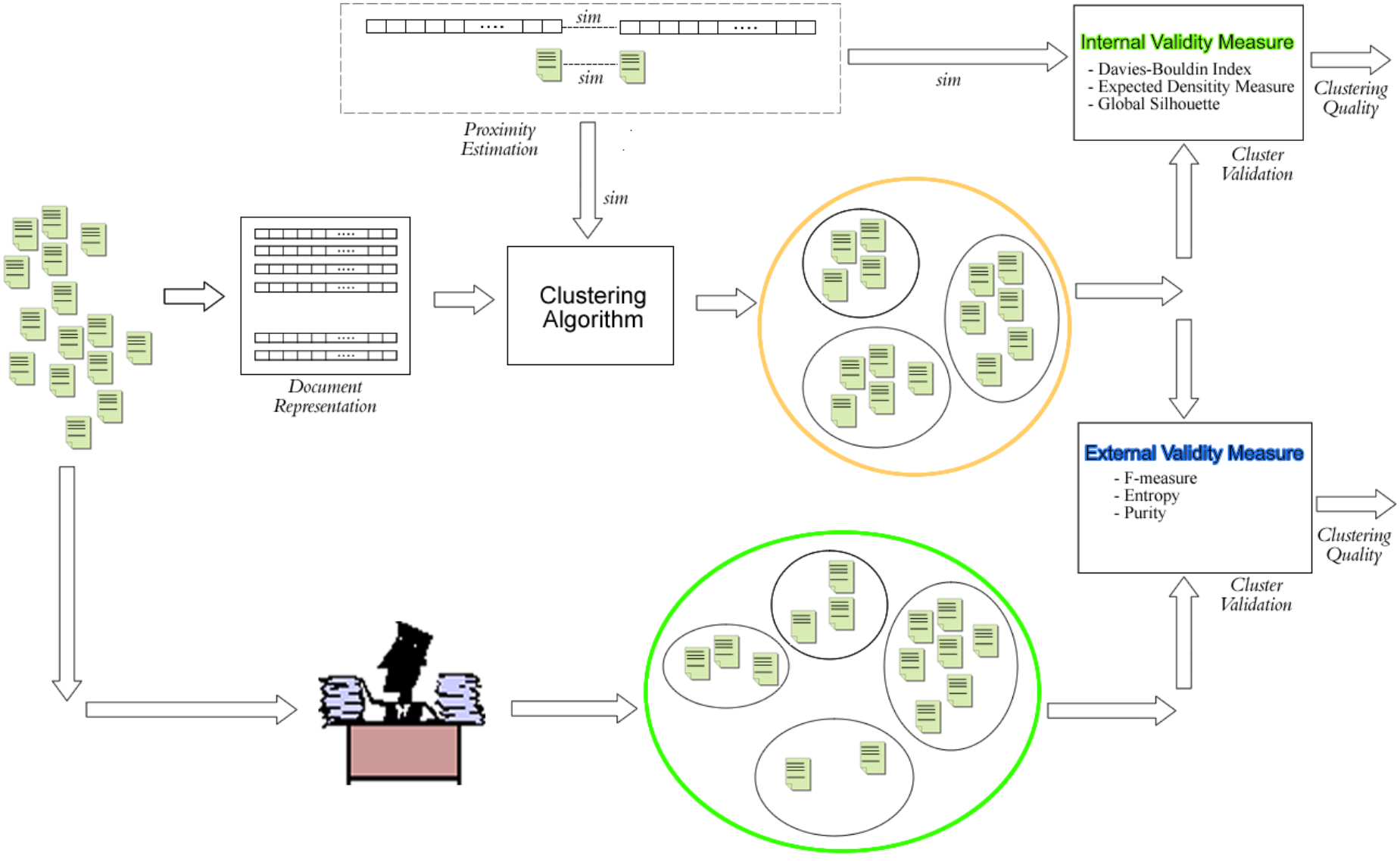


A more detailed look to the document clustering process...



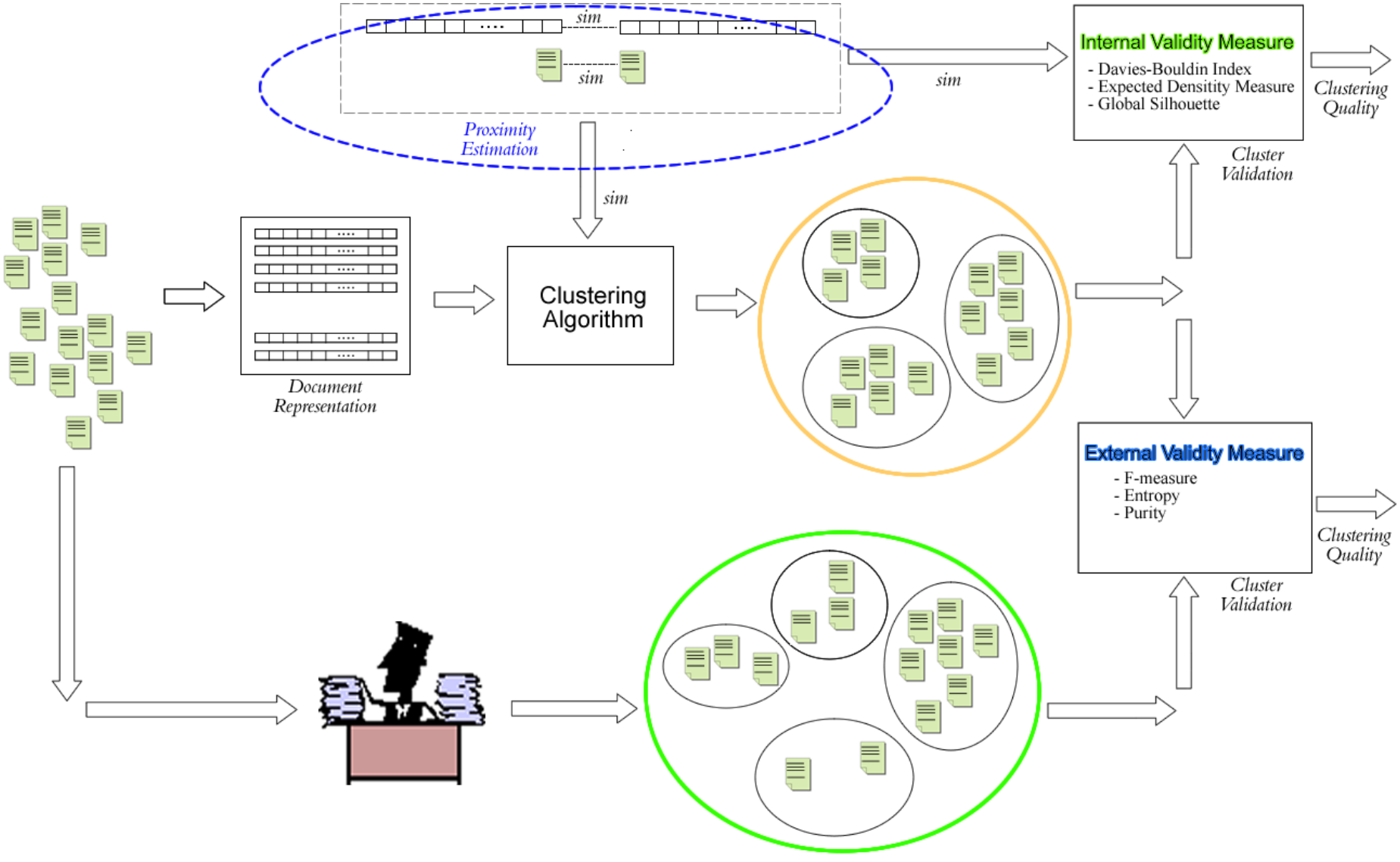
Main ideas behind our approach

First: identify in this process two main components...



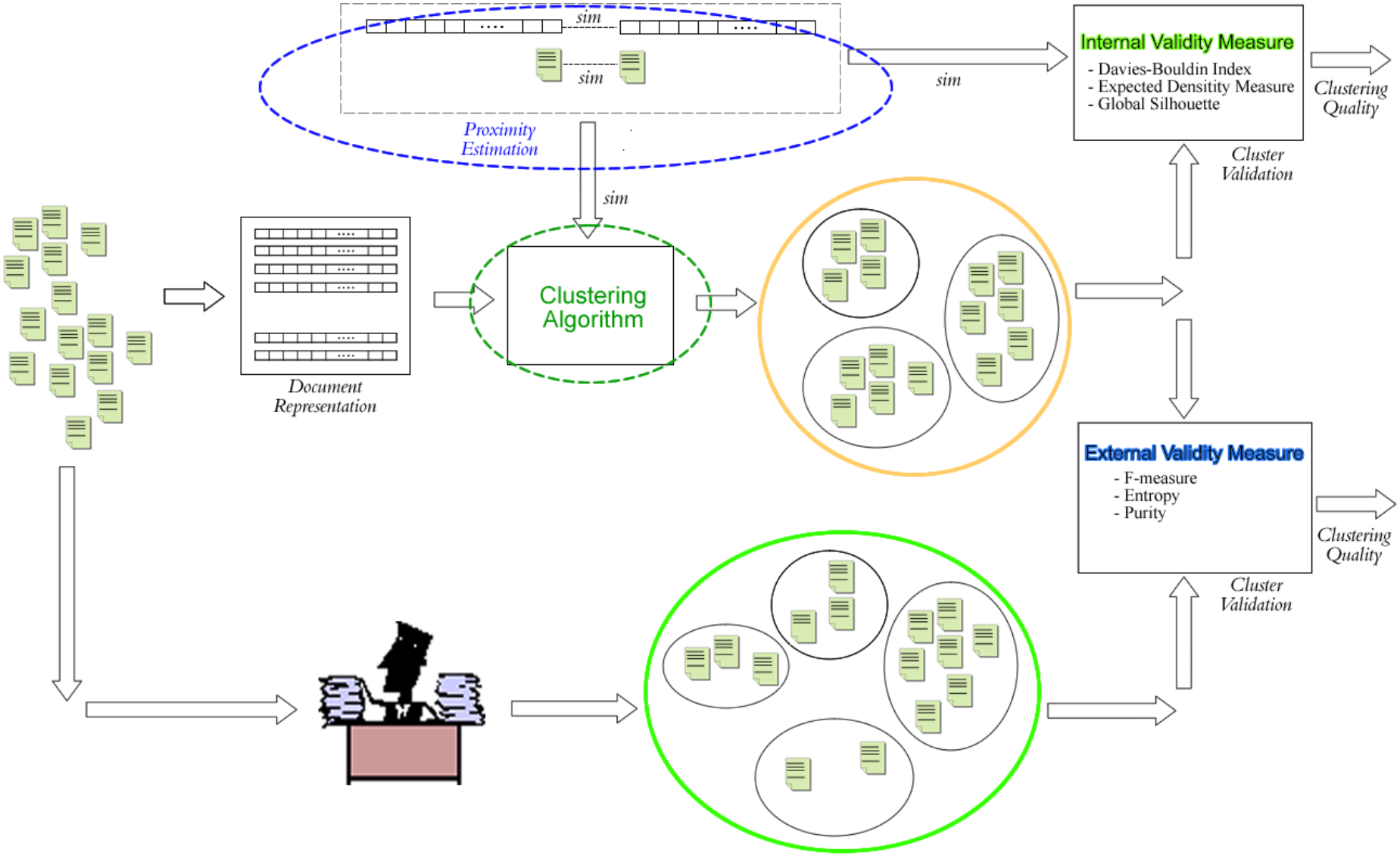
Main ideas behind our approach

the proximity estimation...



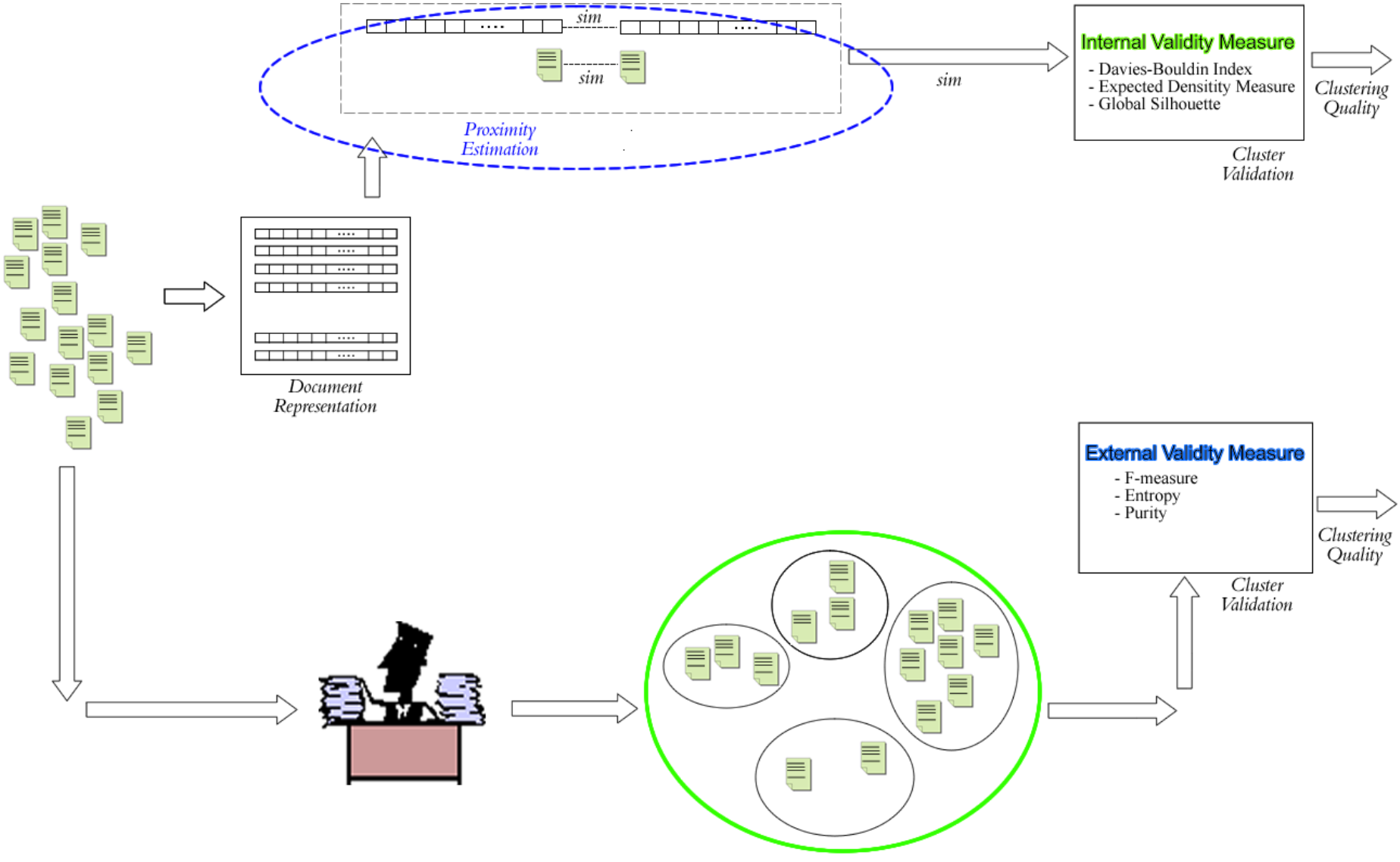
Main ideas behind our approach

the clustering algorithm itself ...



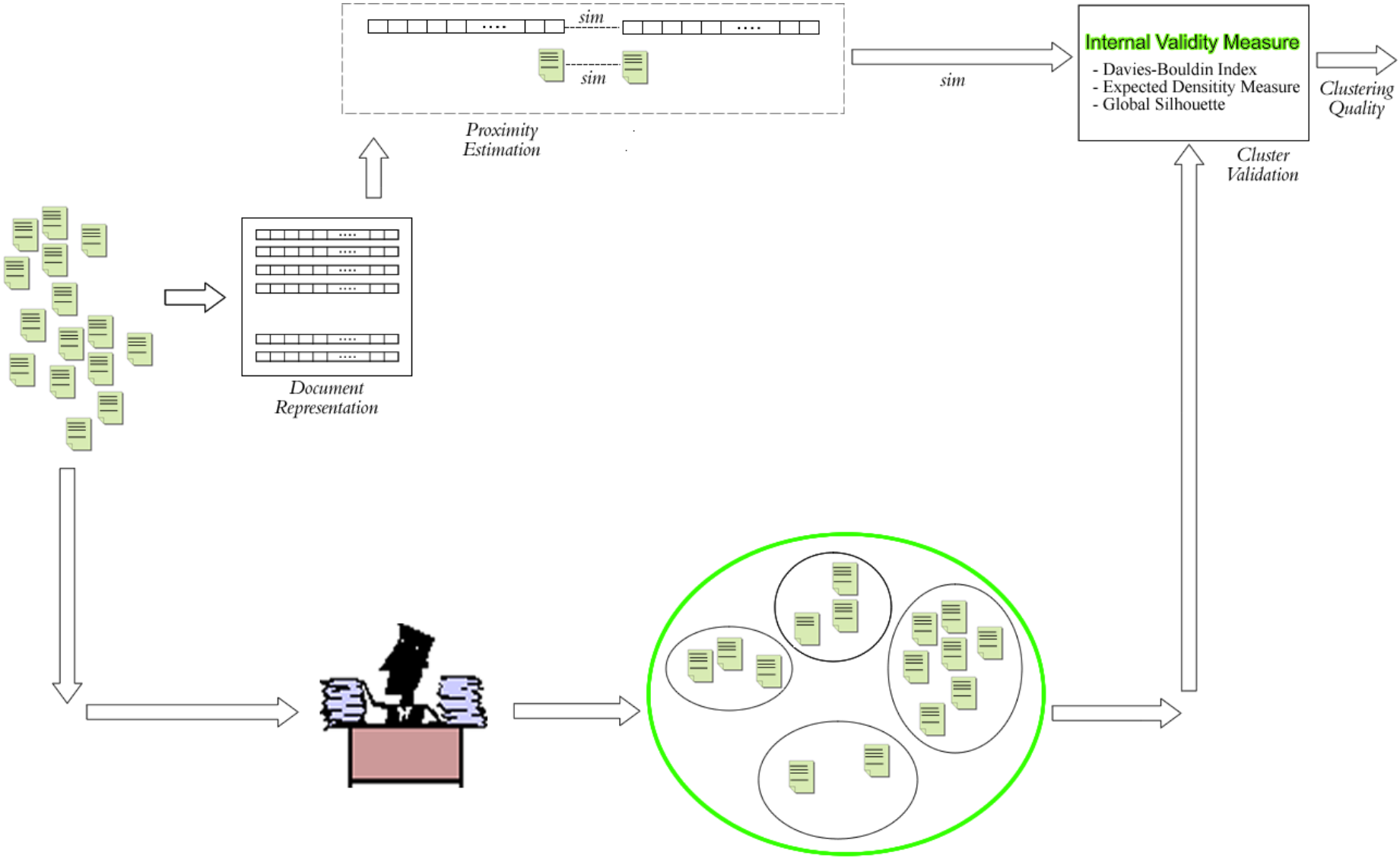
Main ideas behind our approach

Second: concentrate our attention on the proximity estimation



Main ideas behind our approach

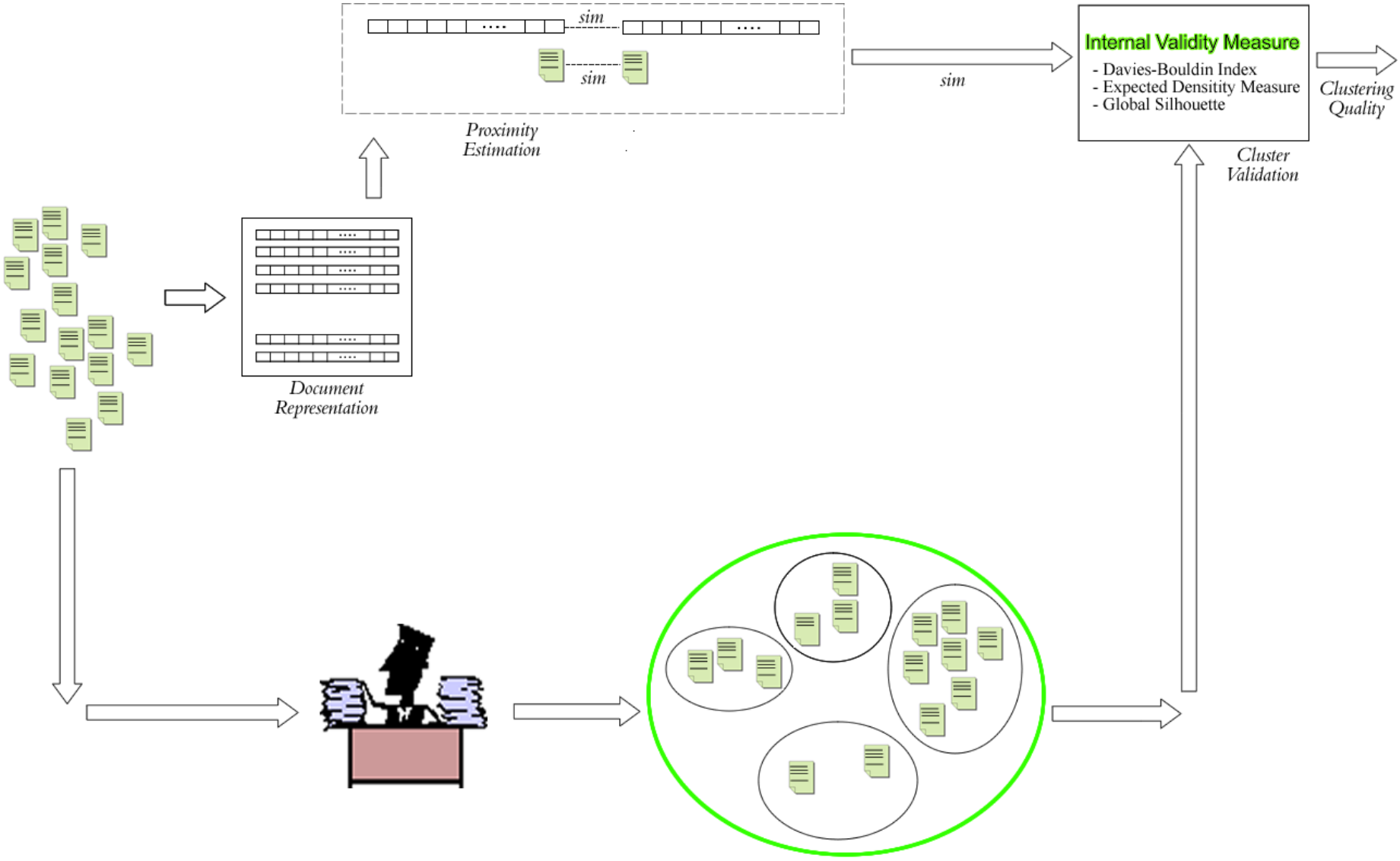
Third: to use validity measures on the "true" categorization



Underlying reasons of this approach

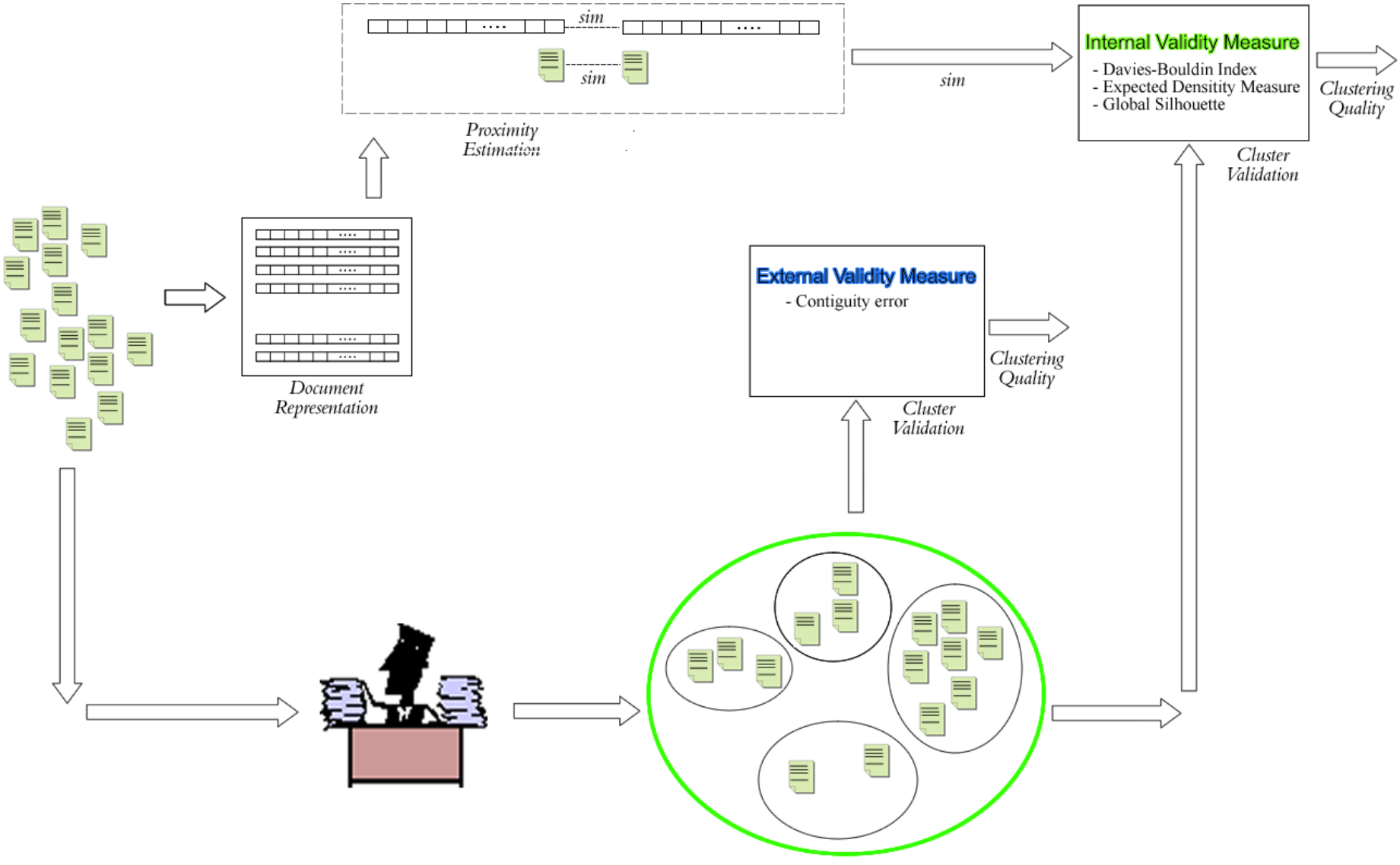
- internal validity measures are usually based on the similarity measure.
- If these measures are not able to detect any interesting structural property when applied to the “true categorization”, this fact can be considered enough evidence that the similarity measure is not adequately expressing the semantic proximity between documents.

Fourth: a new external validity measure



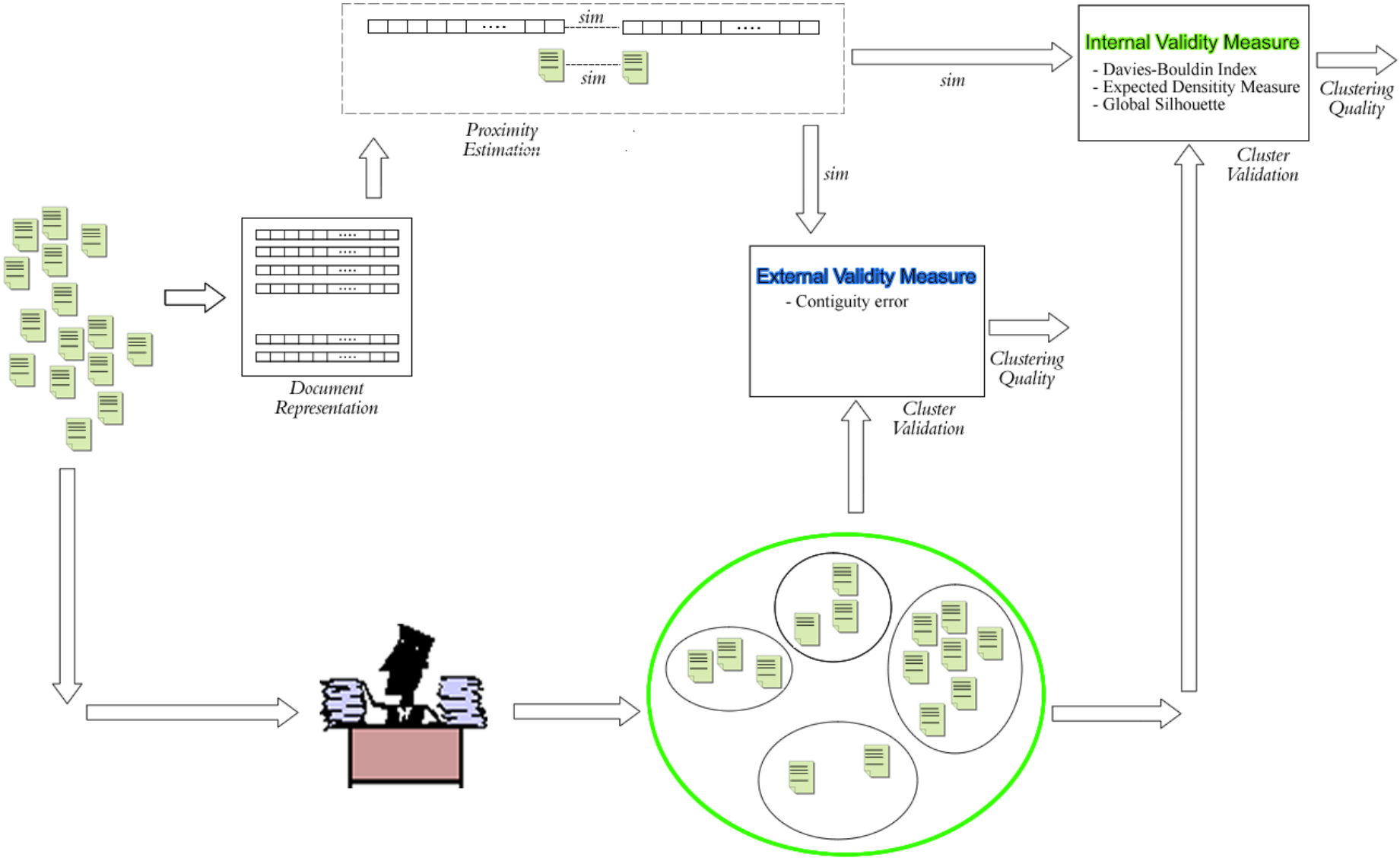
Main ideas behind our approach

Fourth: a new external validity measure, the contiguity error



Main ideas behind our approach

... based on the similarity estimation



The contiguity error

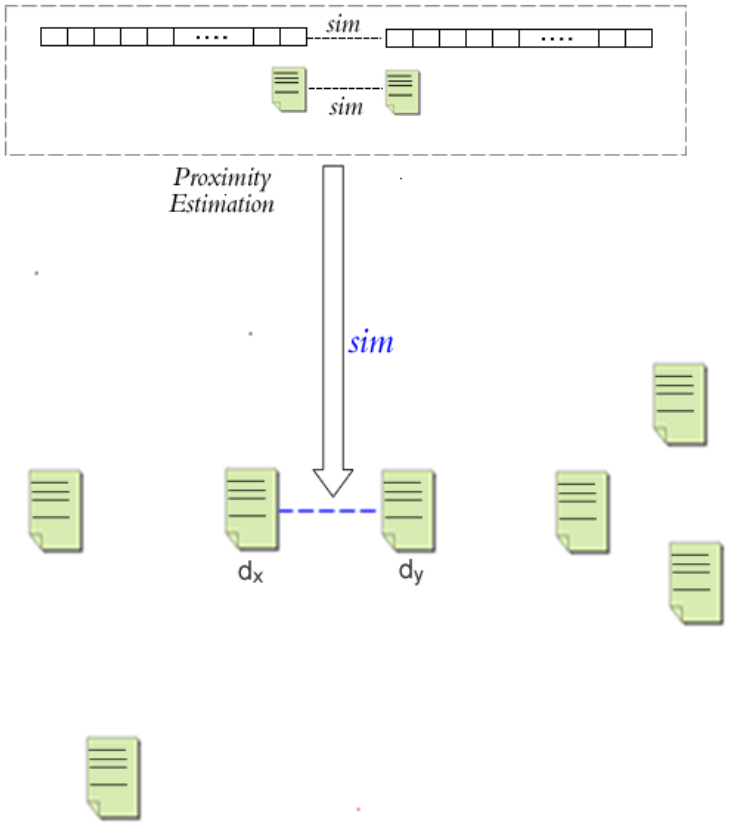
Question: how many contiguity errors a similarity measure produces respect to the clustering specified by the expert?.

The contiguity error

Question: how many contiguity errors a similarity measure produces respect to the clustering specified by the expert?.

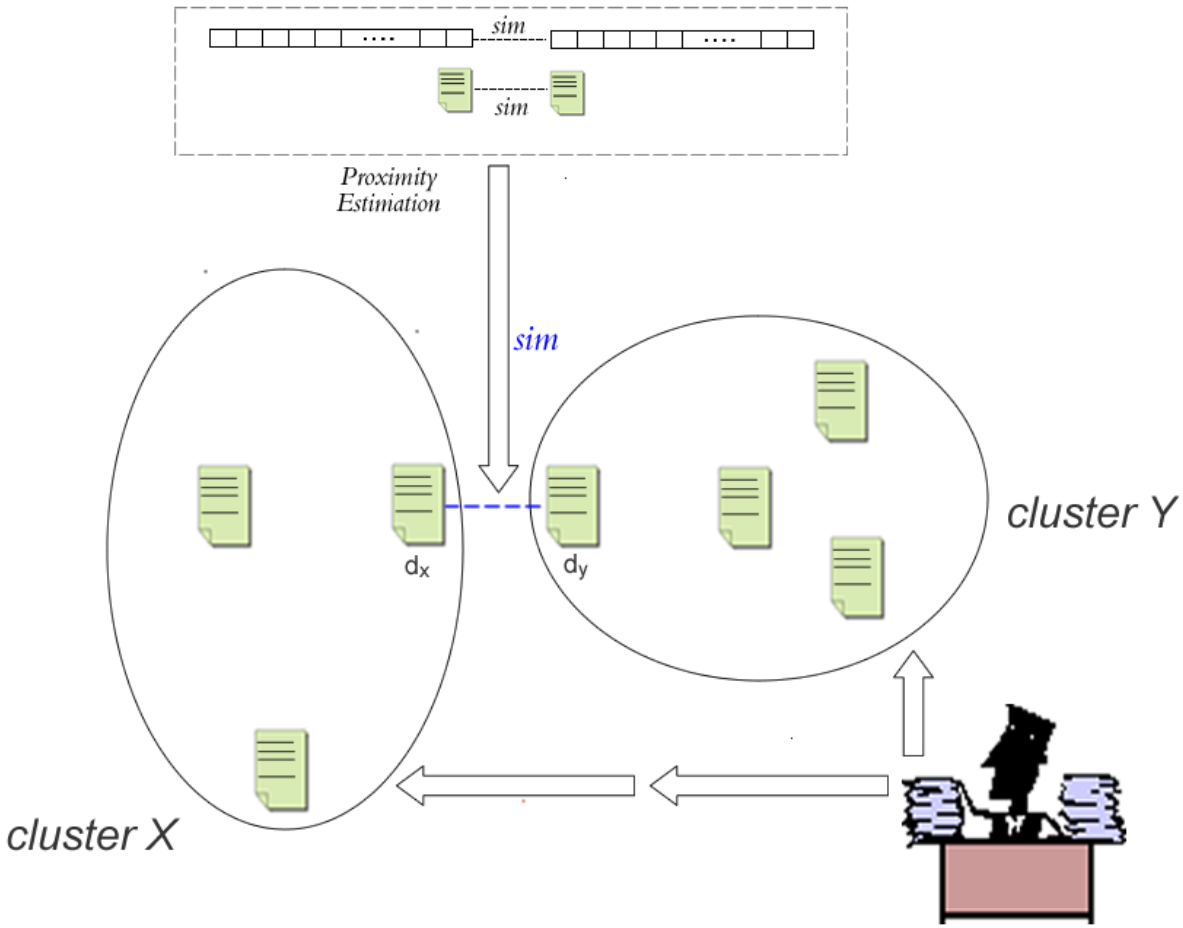
The contiguity error

according to *sim*, d_x has a document d_y as its nearest neighbour, ...



The contiguity error

according to *sim*, d_x has a document d_y as its nearest neighbour, but they were categorized in different clusters!!!

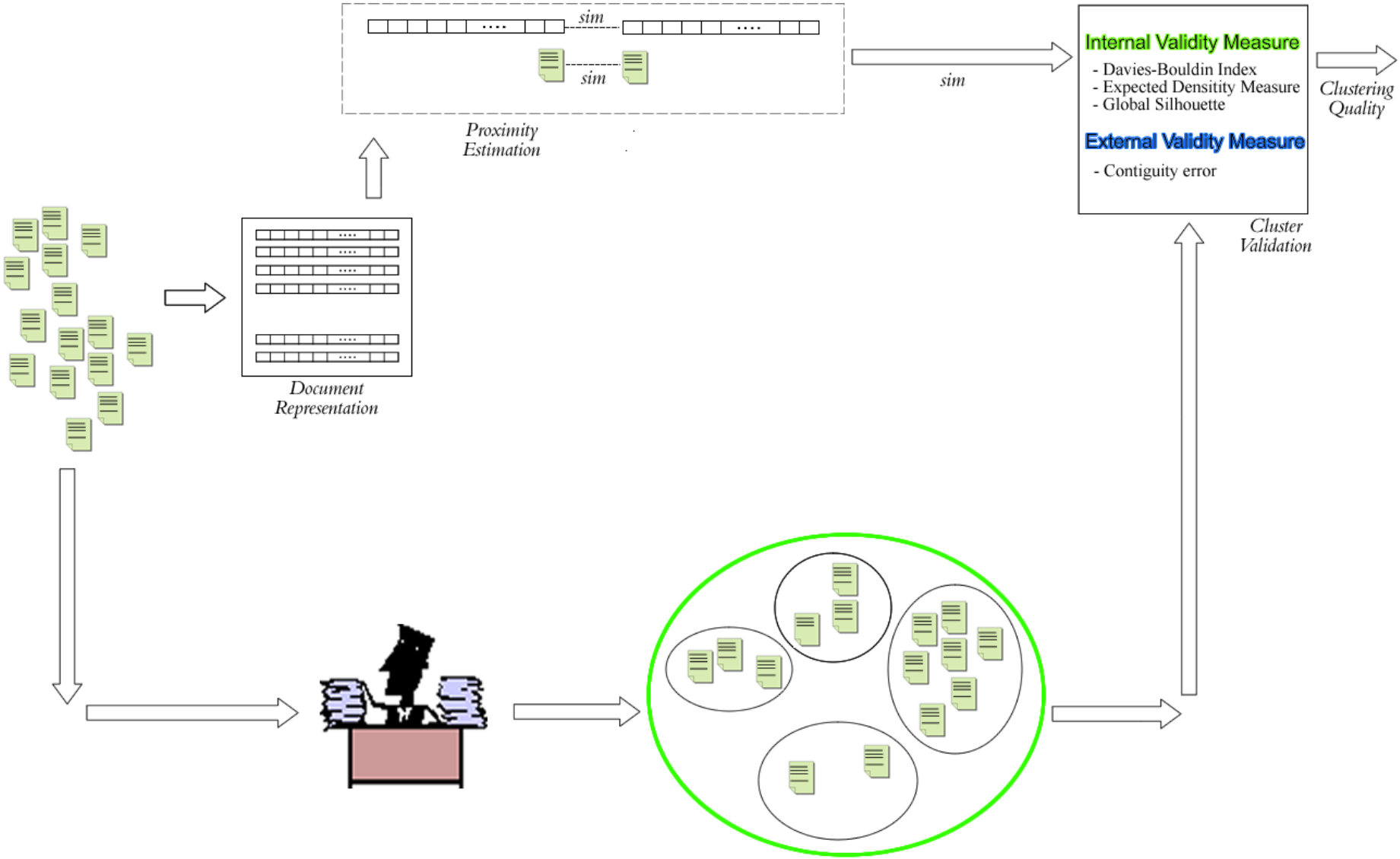


The contiguity error

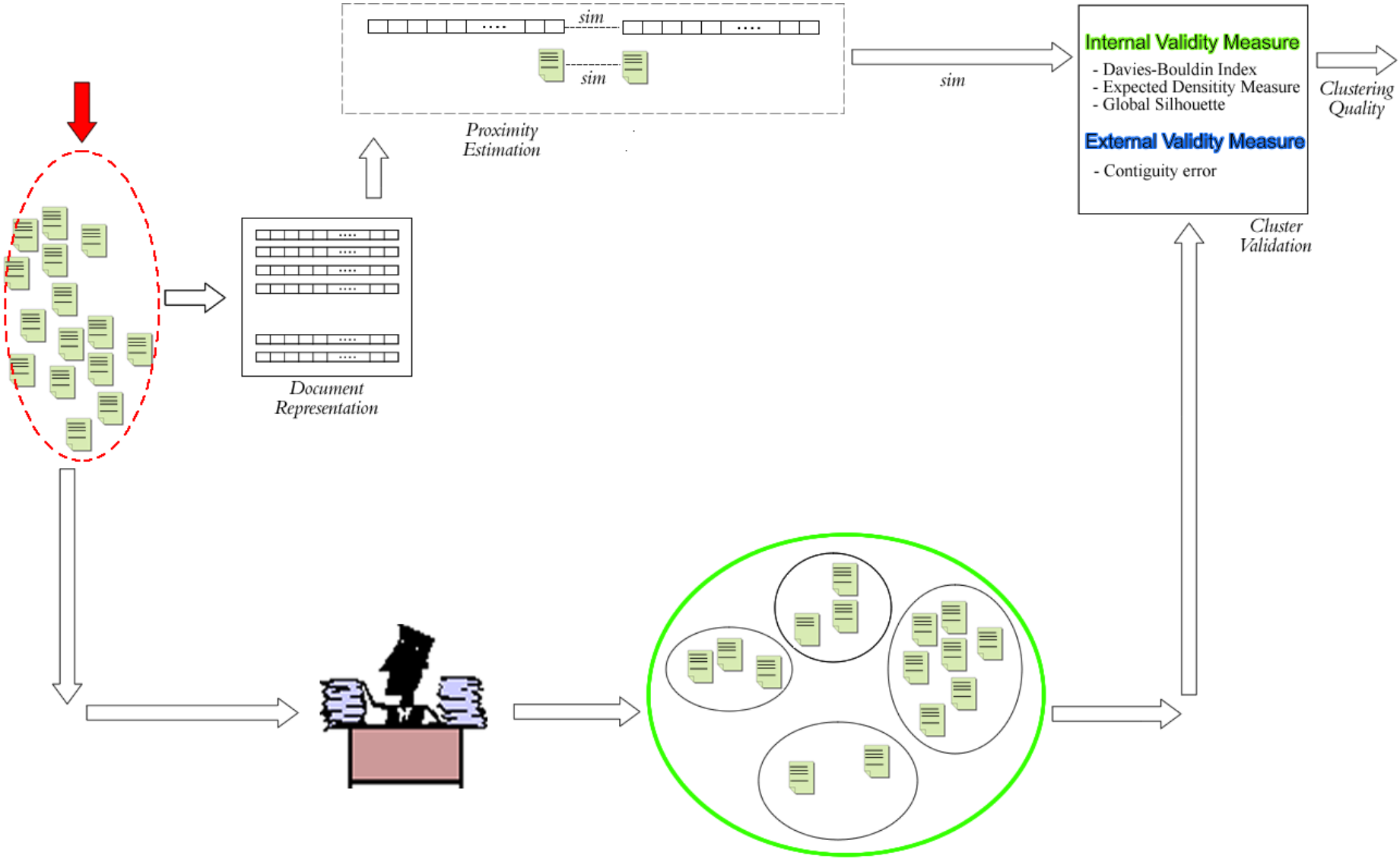
Intuitive idea

The contiguity error (*CE*) of a similarity measure with respect to a collection, is the total number of contiguity errors that this measure commits on all the documents in the collection.

Summary of our approach...



Data Sets


sim
sim
sim
Proximity Estimation
Internal Validity Measure
- Davies-Bouldin Index
- Expected Densitity Measure
- Global Silhouette
External Validity Measure
- Contiguity error
Clustering Quality
Cluster Validation
Document Representation

Criteria for selecting the corpora

1. Data sets with *different complexity level* according to:
   1. Length of documents: short (high) vs. long (low)
   2. How related the topics corresponding to the different groups are: very related (high) vs. little related(low)
2. Small collections with the same number of documents and number of groups

Difficulty of the Corpora

Corpus

Corpus	Terms $\times$ text
Micro4News	2616.95
EasyAbstracts	192.93
CICling-2002	70.45

Difficulty of the Corpora

Corpus

Corpus	Terms × text
Micro4News	2616.95
EasyAbstracts	192.93
CICling-2002	70.45

CICling-2002

1. short documents
2. related topics
3. high complexity

Difficulty of the Corpora

Corpus

Corpus	Terms × text
Micro4News	2616.95
EasyAbstracts	192.93
CICling-2002	70.45

EasyAbstracts

1. short documents
2. topics well differentiated
3. medium complexity

Difficulty of the Corpora

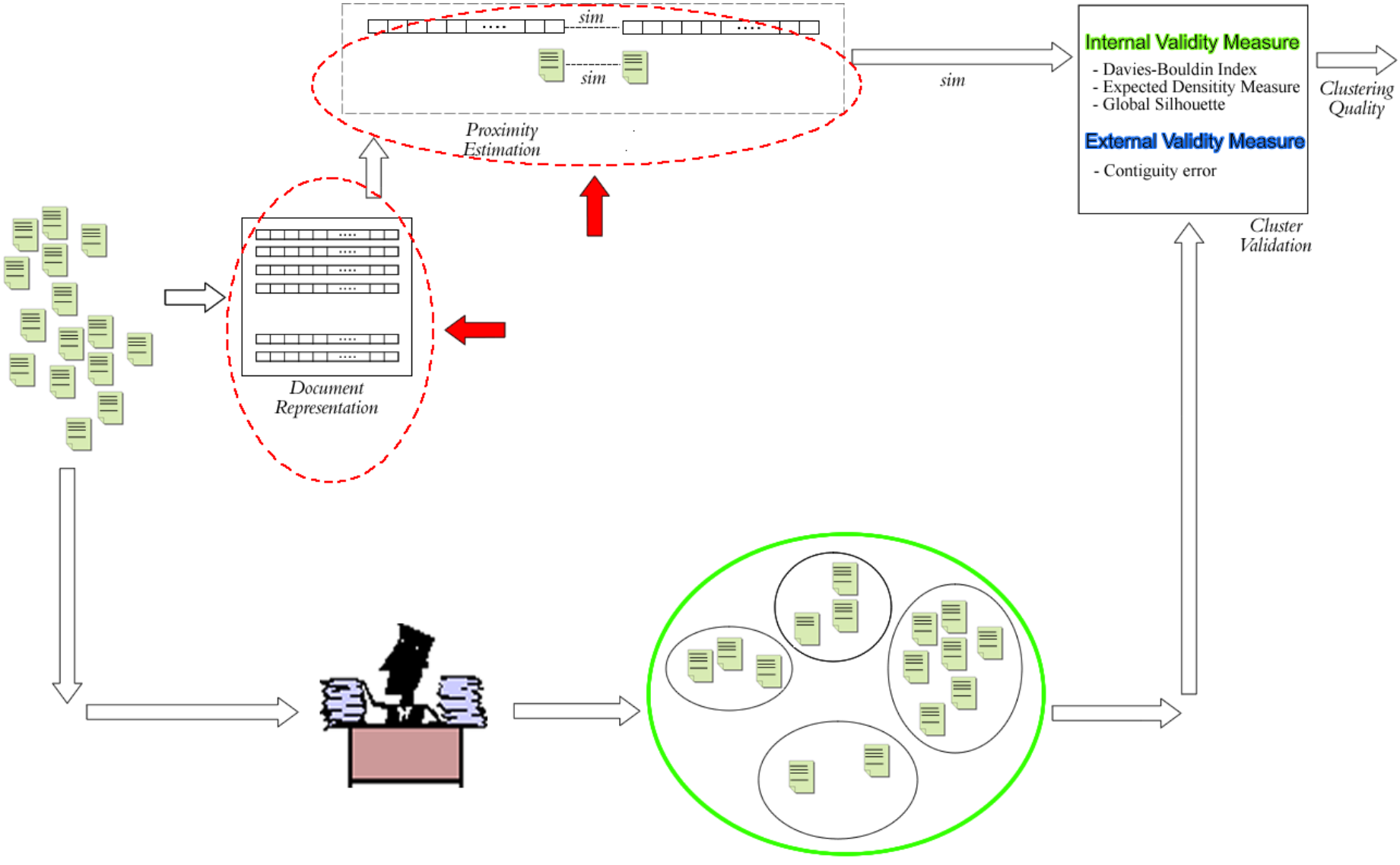
Corpus

Corpus	Terms × text
Micro4News	2616.95
EasyAbstracts	192.93
CICling-2002	70.45

Micro4News

1. long documents
2. topics well differentiated
3. low complexity

Document Representation and similarity estimation



Document Representation and similarity estimation

Some popular alternatives for representing the documents:

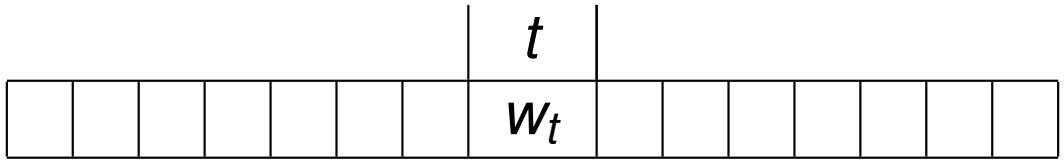
- The Vector Space Model with a family of codification schemes.
- The set model.
- BM-25
- LSI

...and for estimating their similarity:

- Cosine similarity.
- Euclidian distance.
- Jaccard coefficient.

Document Representation

SMART codifications



$$w_t = TF'_{d,t} \cdot IDF'_t \cdot NORM$$

Term Frequency

$n = TF_{d,t}$

$b = 1$

$m = \frac{TF_{d,t}}{max_t(TF_{d,t})}$

$a =$

$0.5 + 0.5\frac{TF_{d,t}}{max_t(TF_{d,t})}$

$l = 1 + \log(TF_{d,t})$

IDF

$n = 1$

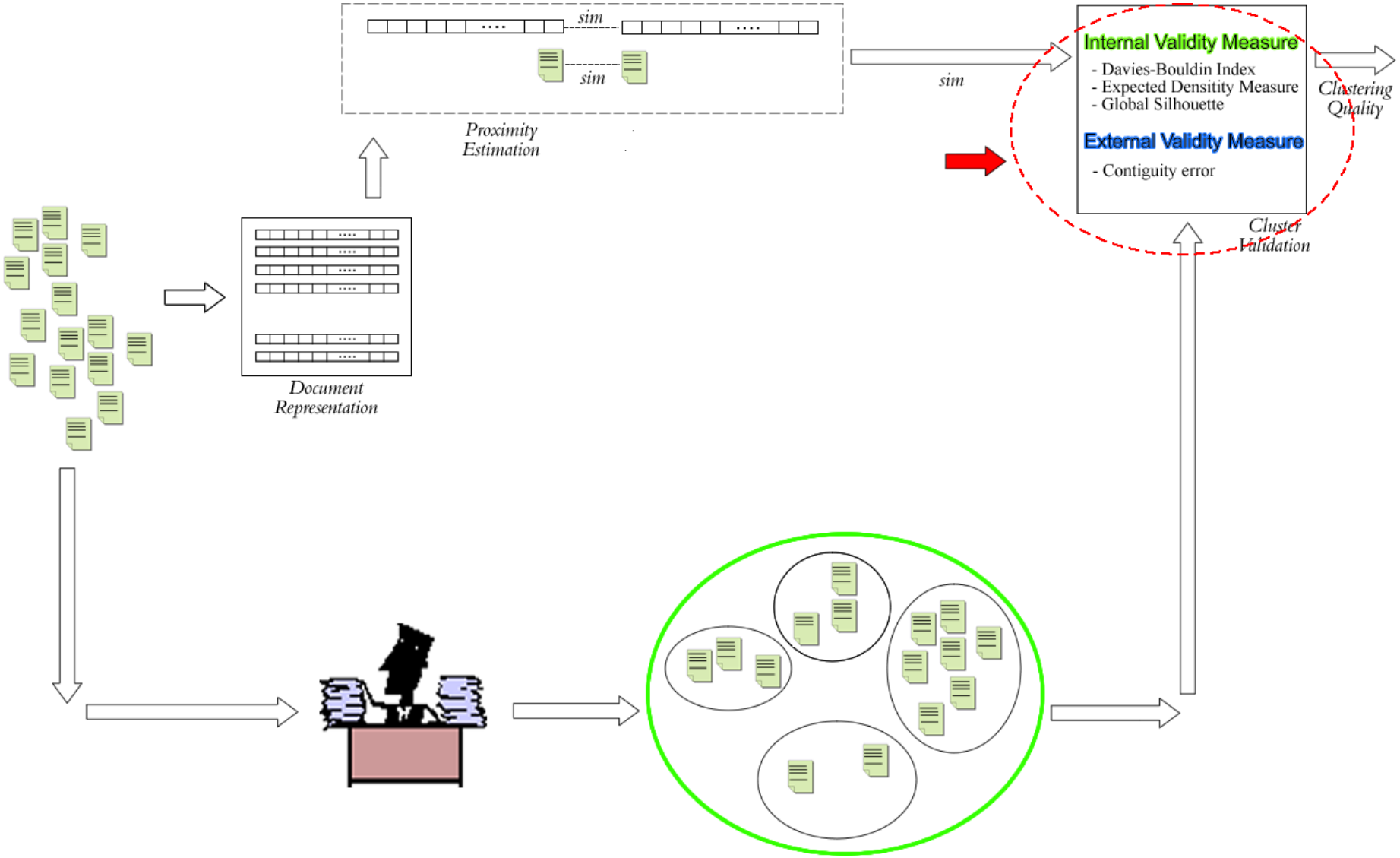
$t = \log(\frac{N}{DF_t})$

NORM

$n = 1$

$c = \frac{1}{\sqrt{\sum_t (TF'_{d,t} IDF'_t)^2}}$

Hardness estimation



Which internal validity measure should we use?

Different internal validity measures attempt to identify specific structural properties of the clusterings like cohesion, separation, density or some combination of these properties.

Which internal validity measure should we use?

Different internal validity measures attempt to identify specific structural properties of the clusterings like cohesion, separation, density or some combination of these properties.

- the *Dunn Index Family*
- the *Davies-Bouldin Index*
- the *Silhouette Coefficient*
- the *Λ-Measure*
- the *Expected Density Measure*

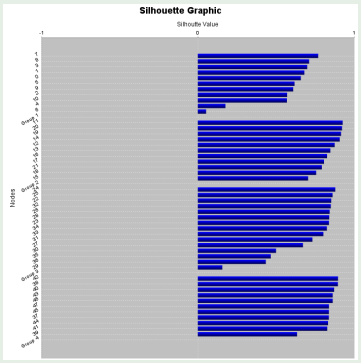
We address this problem avoiding establish a commitment with a particular validity measure and considering a representative group of measures instead.

The Micro4News Corpus

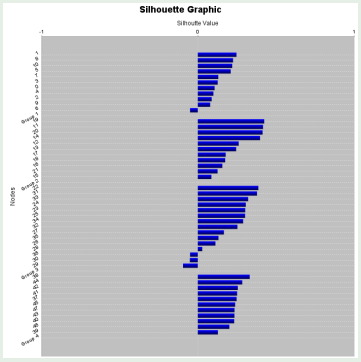
Values of validity measures

Cod.	CE	EDM	DB	Dunn	GS
atc	0	0.9	1.64	0.76	0.46
btc	0	0.9	1.64	0.76	0.46
mtc	**0**	**1.07**	**1.33**	0.76	**0.73**
ntc	**0**	**1.07**	1.34	0.74	**0.73**
Jac	0	0.78	2.10	0.50	0.2
anc	1	0.77	2.48	0.85	0.16
ltc	1	0.92	1.59	0.77	0.50
bnc	1	0.77	2.45	0.85	0.17
lnc	1	0.78	2.52	**0.87**	0.14
mnc	10	0.82	2.89	0.75	0.02
nnc	10	0.82	3.38	0.74	0.02

Silhouette (Cos)



Silhouette (Jac)

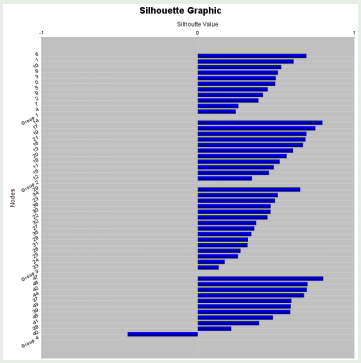


The EasyAbstracts Corpus

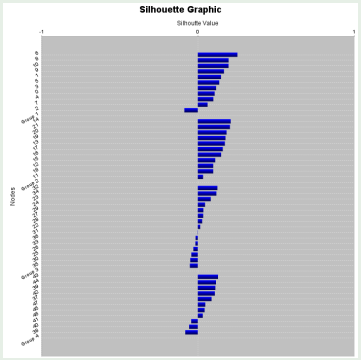
Values of validity measures

Cod.	CE	EDM	DB	Dunn	GS
mtc	**4**	**0.93**	**1.57**	0.71	**0.47**
ntc	**4**	**0.93**	**1.57**	0.71	**0.47**
ltc	5	0.89	1.7	0.71	0.33
atc	5	0.88	1.72	0.71	0.31
btc	6	0.88	1.74	0.71	0.28
lnc	11	0.73	3.57	0.86	0.07
anc	11	0.72	3.49	0.85	0.07
Jac	13	0.74	2.15	0.5	0.08
bnc	15	0.72	3.28	0.82	0.07
mnc	20	0.75	4.91	**0.87**	0.02
nnc	20	0.75	4.91	**0.87**	0.02

Silhouette (Cos)



Silhouette (Jac)

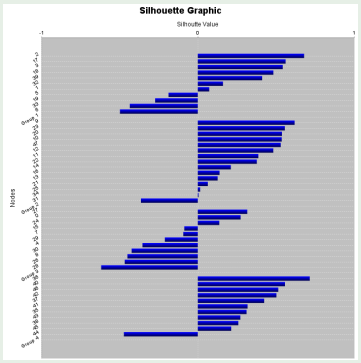


The CICling-2002 Corpus

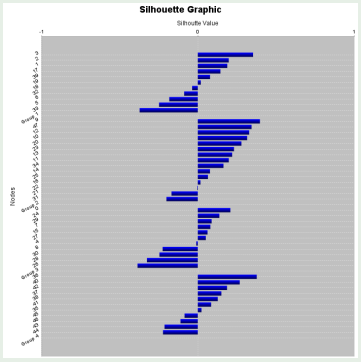
Values of validity measures

Cod.	CE	EDM	DB	Dunn	GS
mnc	**16**	0.8	2.21	0.79	**0.15**
nnc	**16**	0.8	2.21	0.79	**0.15**
btc	18	0.84	1.82	0.74	0.07
anc	21	0.76	2.45	**0.8**	0.07
Jac	22	0.79	2.28	0.53	0.05
atc	22	0.85	1.8	0.74	0.1
bnc	22	0.75	2.51	**0.8**	0.04
ltc	23	0.85	1.8	0.74	0.1
lnc	23	0.76	2.45	**0.8**	0.08
mtc	23	**0.87**	**1.76**	0.74	**0.15**
ntc	23	**0.87**	**1.76**	0.74	**0.15**

Silhouette (Cos)



Silhouette (Jac)



Conclusions

- our approach can be an interesting tool for determining the hardness of corpora used as testbed in clustering of short-text corpora.
- traditional methods for computing similarity measures can be used with short-text corpora with well differentiated topics but more elaborated approaches are required for obtain acceptable results with narrow domain short-text corpora.
- Silhouette Global, Expected Density Measure and Contiguity Error exhibit an interesting consistency level in all the collections considered and seem to be the most informative for determining the most adequate similarity scheme for each corpus

Future work

- To extend our work to other corpora
- To use other more elaborated document representation approaches.
- To investigate how robust the different clustering algorithms are to the different error levels exhibited by the similarity measures.
- To use semi-supervised clustering approaches that automatically adapt the similarity estimation
- To use the best internal validity measures as objective functions to be optimized.

Questions?

Questions?

Thank You very much for your attention...

Micro4News Description

Distribution of documents

Category	# docs
windows.misc	11
sci.med	15
rec.autos	11
soc.religion.med	11

1. Long documents
2. Topics Well differentiated
3. Low complexity

Main Characteristics

Feature	Value
Corpus size	722492
# categories	4
# tot. docs	48
# tot. terms	125614
Voc. size	12785
Term per doc.	2616,95
Overl. voc.	0,16

EasyAbstracts Description

Distribution of documents

Category	# docs
Heuristics in Optimization	11
Machine Learning	15
Automated Reasoning	11
Aut. Intelligent Agents	11

1. Short documents
2. Topics Well differentiated
3. Medium complexity

Main Characteristics

Feature	Value
Corpus size	63018
# categories	4
# tot. docs	48
# tot. terms	9261
Voc. size	2169
Term per doc.	192,93
Overl. voc.	0,13

CICling-2002 Description

Distribution of documents

Category	# docs
Linguistics	11
Ambiguity	15
Lexicon	11
Text Processing	11

1. Short documents
2. Related topics
3. High complexity

Main Characteristics

Feature	Value
Corpus size	23971
# categories	4
# tot. docs	48
# tot. terms	3382
Voc. size	953
Term per doc.	70.45
Overl. voc.	0,22

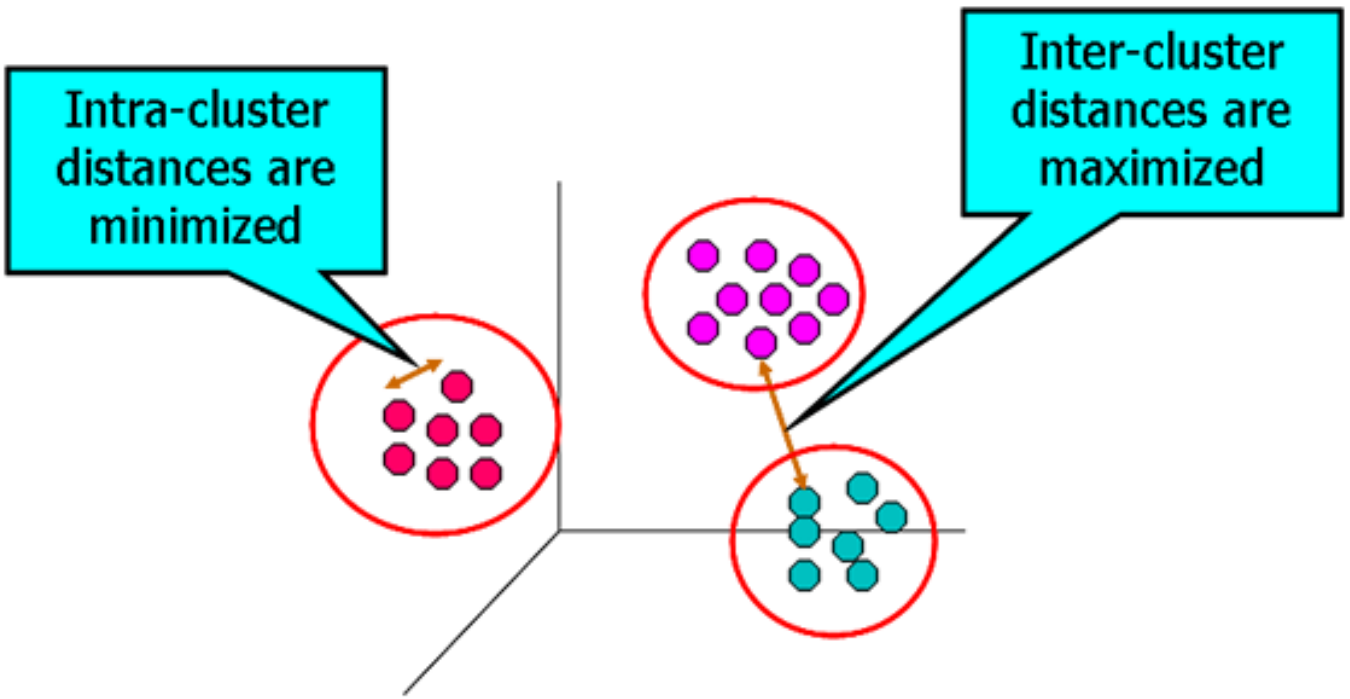
Difficulty of the Corpora

Corpus

Corpus	Terms $\times$ text	Vocab. overlapping
Micro4News	2616.95	0.16
EasyAbstracts	192.93	0.13
CICling-2002	70.45	0.22

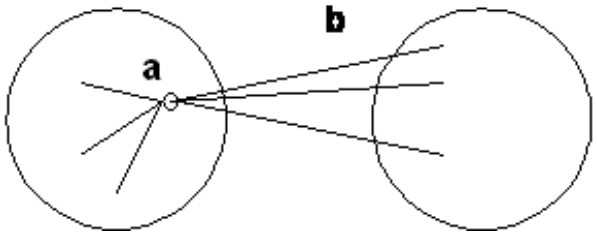
What is Document Clustering?

- Finding groups of documents such that the documents in a group will be similar (or related) to one another and different from (or unrelated to) the documents in other groups



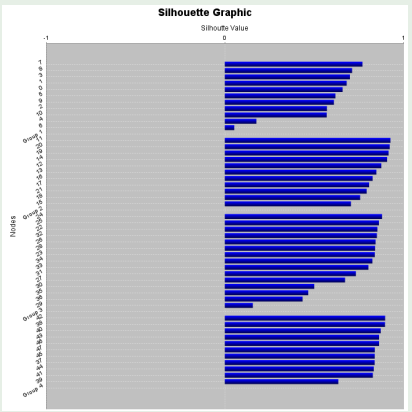
An informative validity measure: Silhouette Coefficient

Combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings



Can calculate the Silhouette width for a cluster or a clustering

Good Clustering



Bad Clustering