Labelling hierarchical clusters of scientific articles

Irina Peganova Yaroslav Nedumov Alena Rebrova

ISP RAS

September 12, 2019

Relevance

Clustering and **labelling clusters** are useful tools to ease the search of scientific articles.

Our aim was develop a method for labelling clusters in Scinoon system.

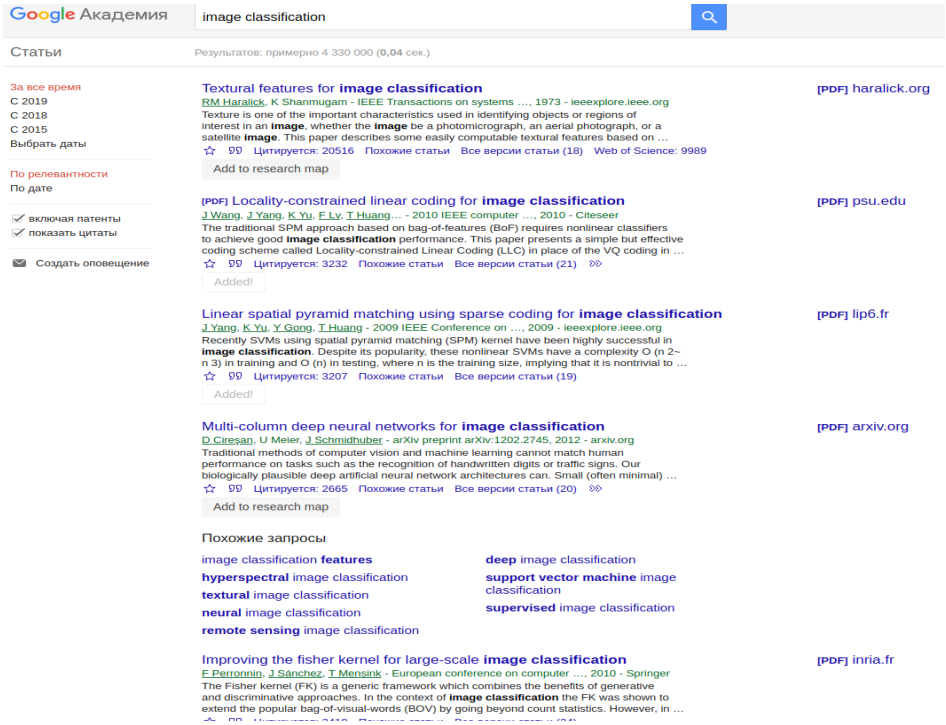


Figure 1: Common search (Google Scholar)

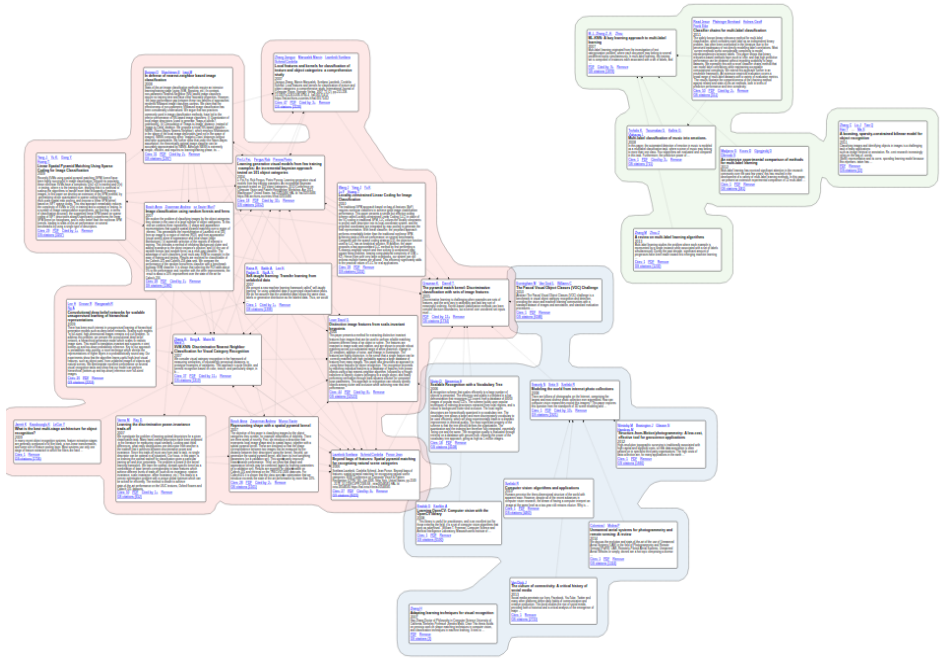


Figure 2: Manually clustered articles (Scinoon)

Our domain

Our domain is **collections of scientific articles** that are:

- quiet specific;
- not large-scale (up to 100);
- represented with their abstracts and meta-data.

Employ hierarchical clustering

Scientific domain is hierarchical, so we decided to label **hierarchical clustering**. Hierarchical clustering lies in building a tree in which a parent cluster consists of its child clusters.

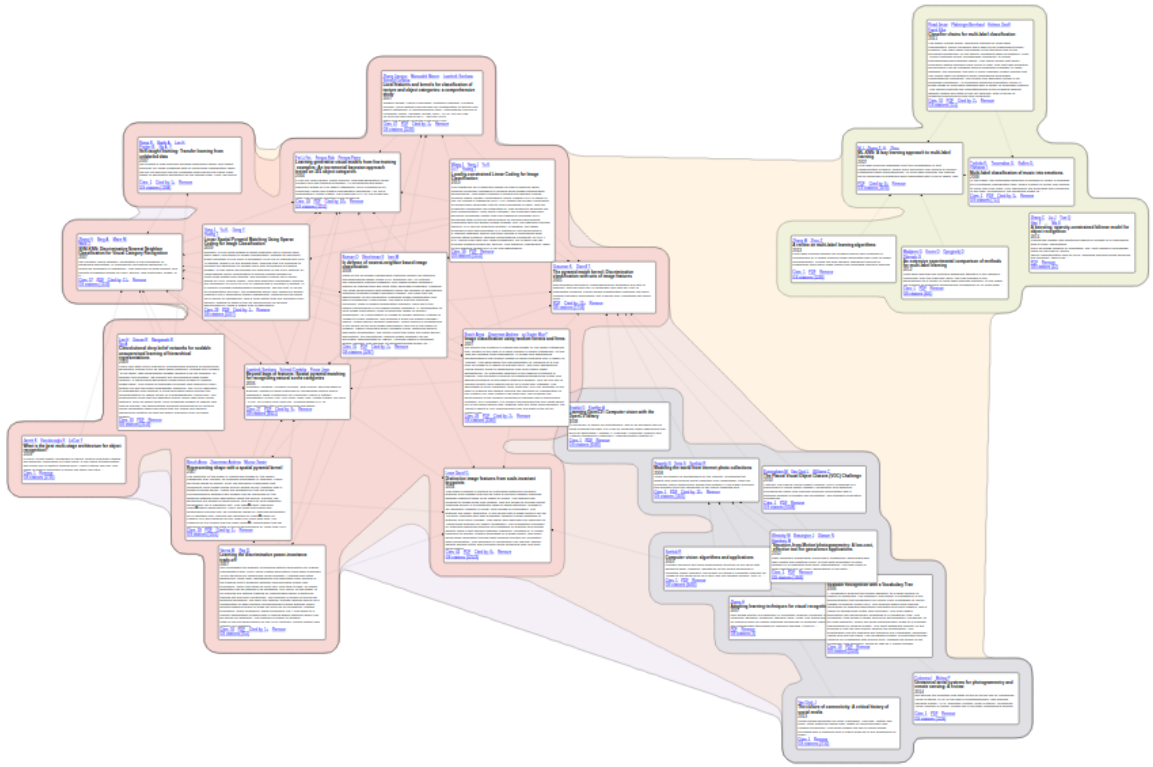


Figure 3: Manually hierarchically clustered articles (Scinoon)

Requirements for clustering descriptions

The basic requirements for clustering descriptions are following (Zhang et al (2009)):

- Conciseness
- Comprehensibility
- Accuracy
- Distinctiveness

Existed solutions: overview

Sources of labels	**Approaches**	**Disadvantages**
External resources	Hyperonyms (WordNet)	The suitable classification either doesn't exist, or lefts behind
	Articles' titles (Wikipedia)	
	Category titles (Open Directory Project)	
Cluster's documents	TF-IDF, TF-ICF like	Level of specificity of label is regulated implicitly (not good for small and quite specific collections)
	Reference-based (χ^2-test, JSD etc)	
	Combined	

Our solution: build on ComboBasic

A plain term extraction algorithm (Astrakhantsev, 2016)

$$ComboBasic(t) = |t| \cdot \log f(t) + \alpha \cdot e_t + \beta \cdot e'_t$$

Allows to customize the level specificity of terms **explicitly** with α and β.

- $|t|$ is the length of t in words
- $f(t)$ is the frequency of t
- e_t is count of longer term candidates (superterms)
- e'_t is count of shorter term candidates (subterms)

Examples

"In this paper, we propose a new method HCBasic for **labelling hierarchical clusters**."

- **"hierarchical clusters"** is more specific than **"clusters"**
- **"labelling"** is more general than **"labelling hierarchical clusters"**

HCBasic: adjusting for hierarchy

The weighting scheme

$$HCBasic(t) = |t| \cdot \log f(t) + \tilde{\alpha} \cdot e_t + \tilde{\beta} \cdot e'_t + \tau(t) + \tilde{\gamma} \cdot \pi(t)$$

- $\tilde{\alpha} = \alpha - 0.1 \cdot pos$
- $\tilde{\beta} = \beta + 0.1 \cdot pos$
- $pos = \dfrac{depth(cluster)}{depth(cluster) + heigth(cluster)}$

HCBasic: benefits from papers' titles

The weighting scheme

$$HCBasic(t) = |t| \cdot \log f(t) + \tilde{\alpha} \cdot e_t + \tilde{\beta} \cdot e'_t + \tau(t) + \tilde{\gamma} \cdot \pi(t)$$

- $\tau(t)$ is the number of articles, in whose titles term t has occurred, normalized with the cluster size

HCBasic: benefits from abstracts' claims

The weighting scheme

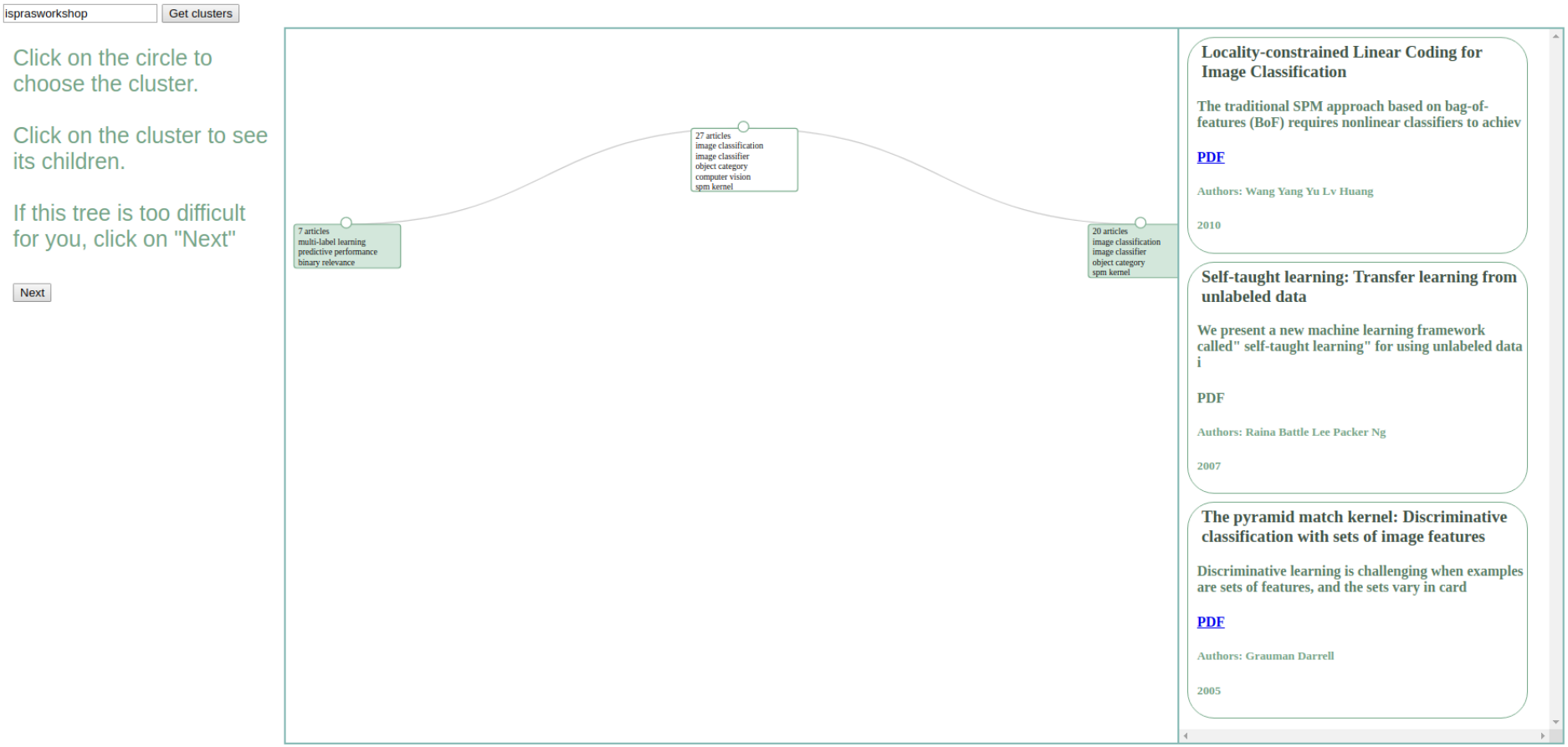
$$HCBasic(t) = |t| \cdot \log f(t) + \tilde{\alpha} \cdot e_t + \tilde{\beta} \cdot e'_t + \tau(t) + \tilde{\gamma} \cdot \pi(t) \quad (1)$$

- $\pi(t)$ is the number of occurrences of a term t in "claim sentences" normalized with its total occurrences.
- $\tilde{\gamma} = 1 + pos$
- $pos = \dfrac{depth(cluster)}{depth(cluster) + heigth(cluster)}$

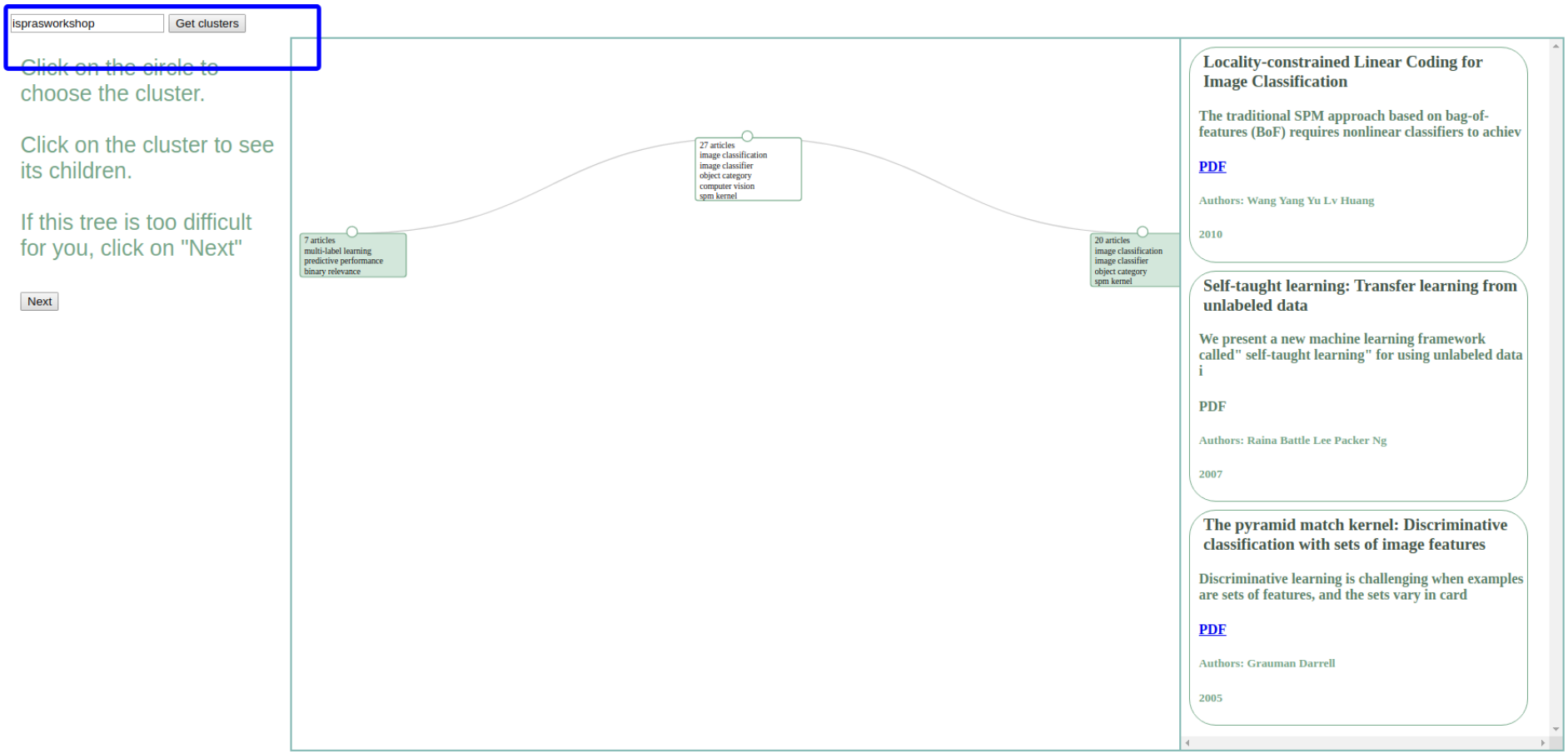
Examples

- In this paper, we propose a new method HCBasic for labelling hierarchical clusters.
- The main contribution of this article is the idea of customizing the level of labels' specificity explicitly.

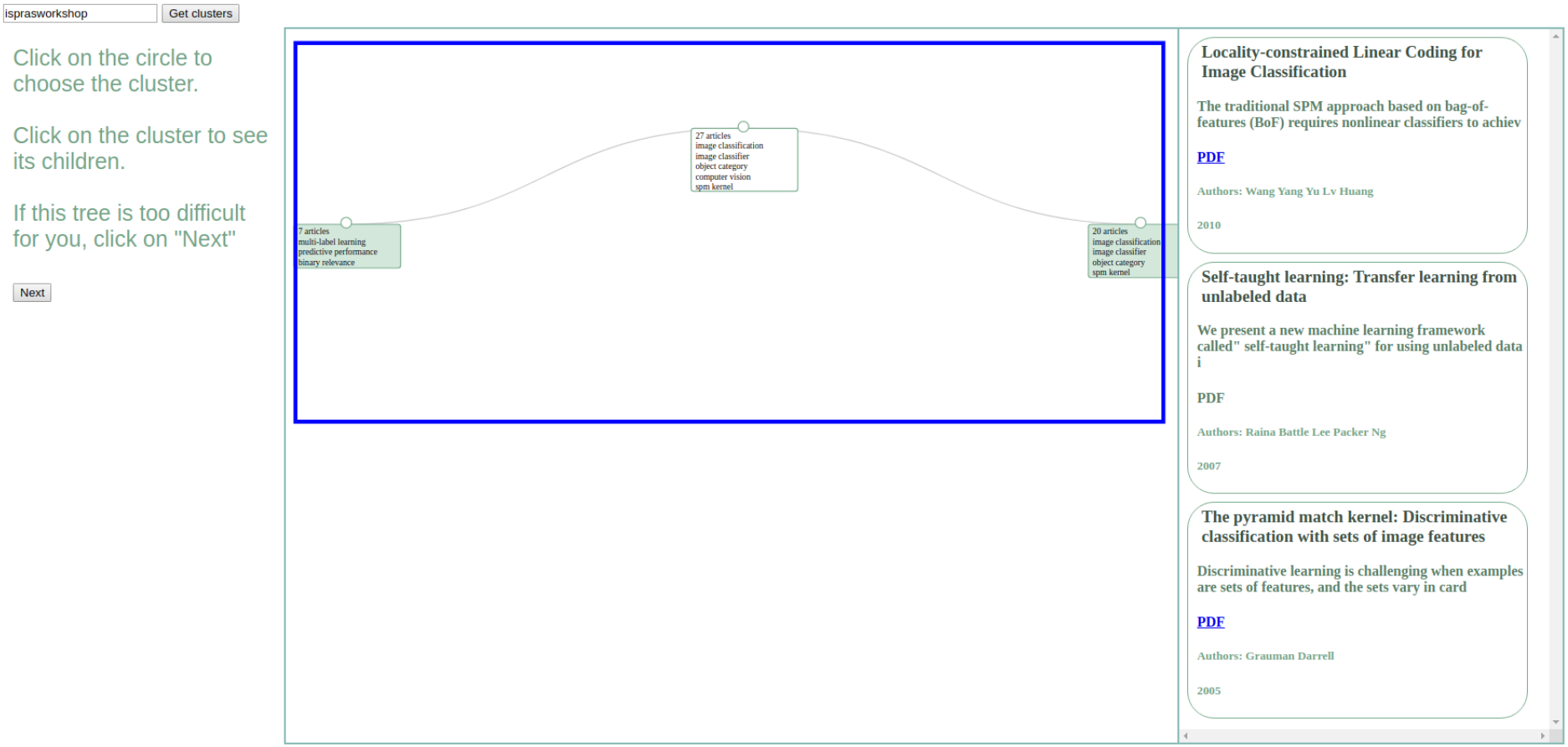
User interface of estimation system (1)



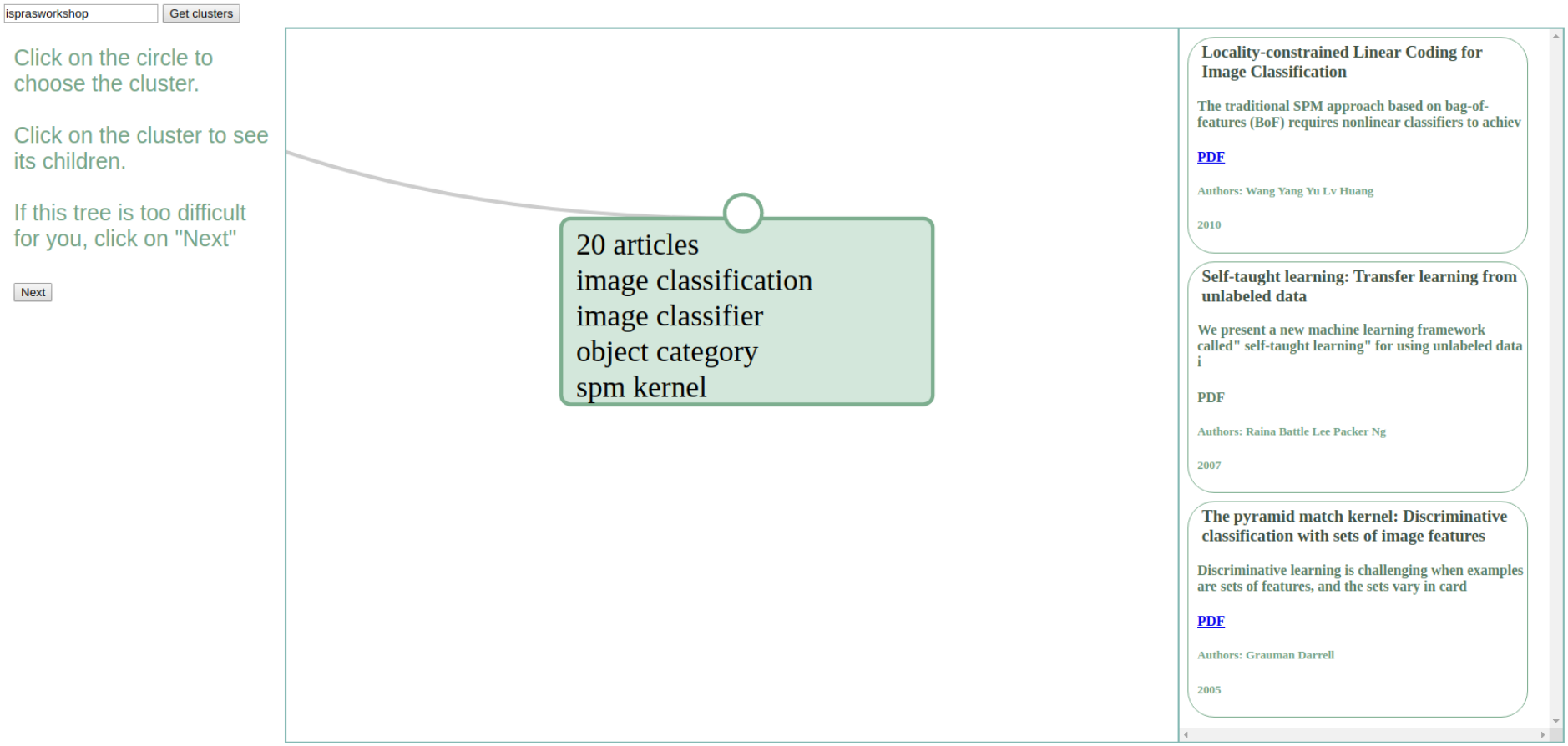
User interface of estimation system (2)



User interface of estimation system: labelled cluster tree

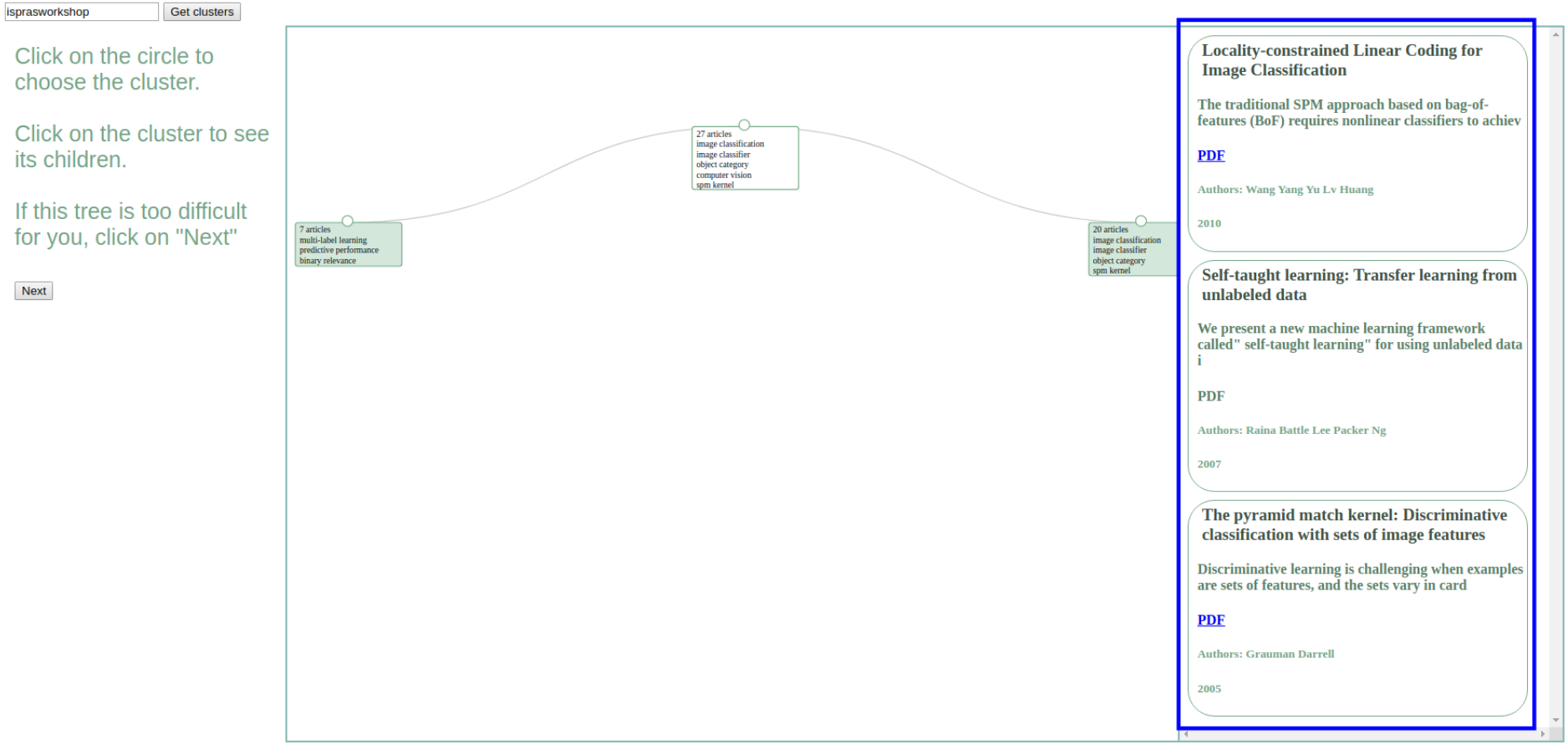


User interface of estimation system: cluster block



User interface of estimation system: asked articles

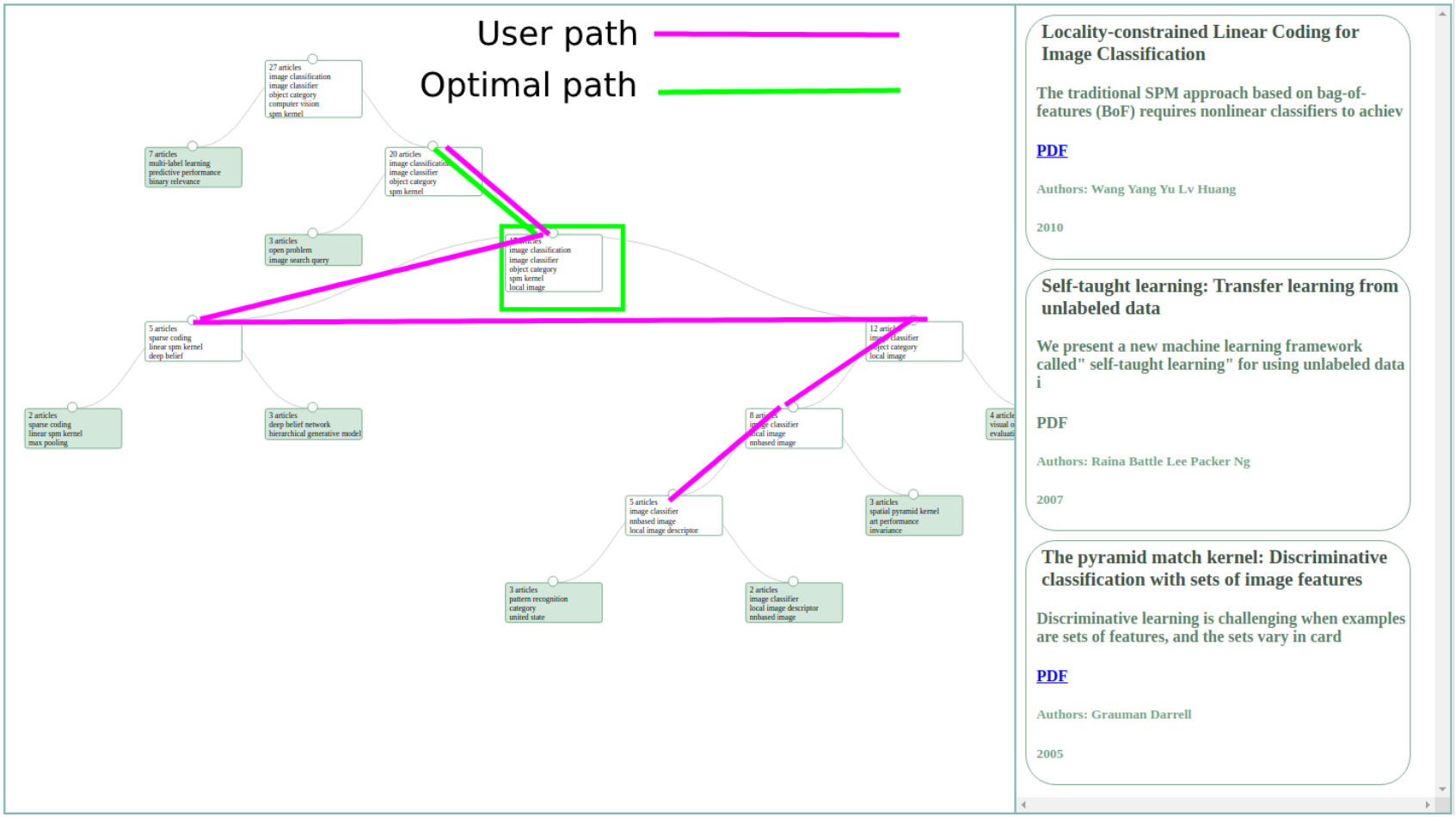
A user was asked to find the less redundant cluster which contains side panel articles.



Benchmarks: PathRatio

How long had the user been searching?

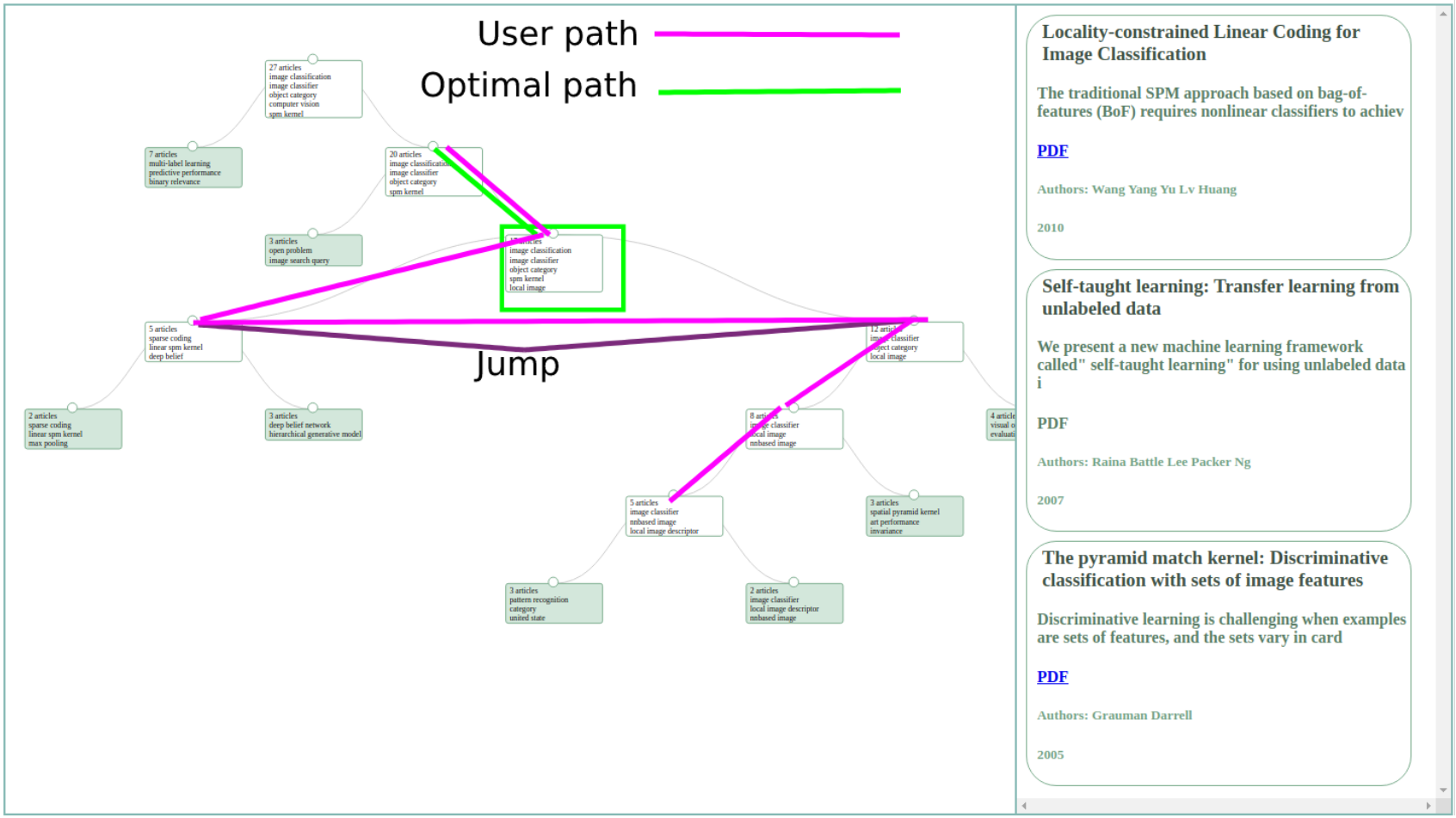
We compute the ratio between the user path and the optimal path.



Benchmarks: “Jumps”

How accurate the labels are?

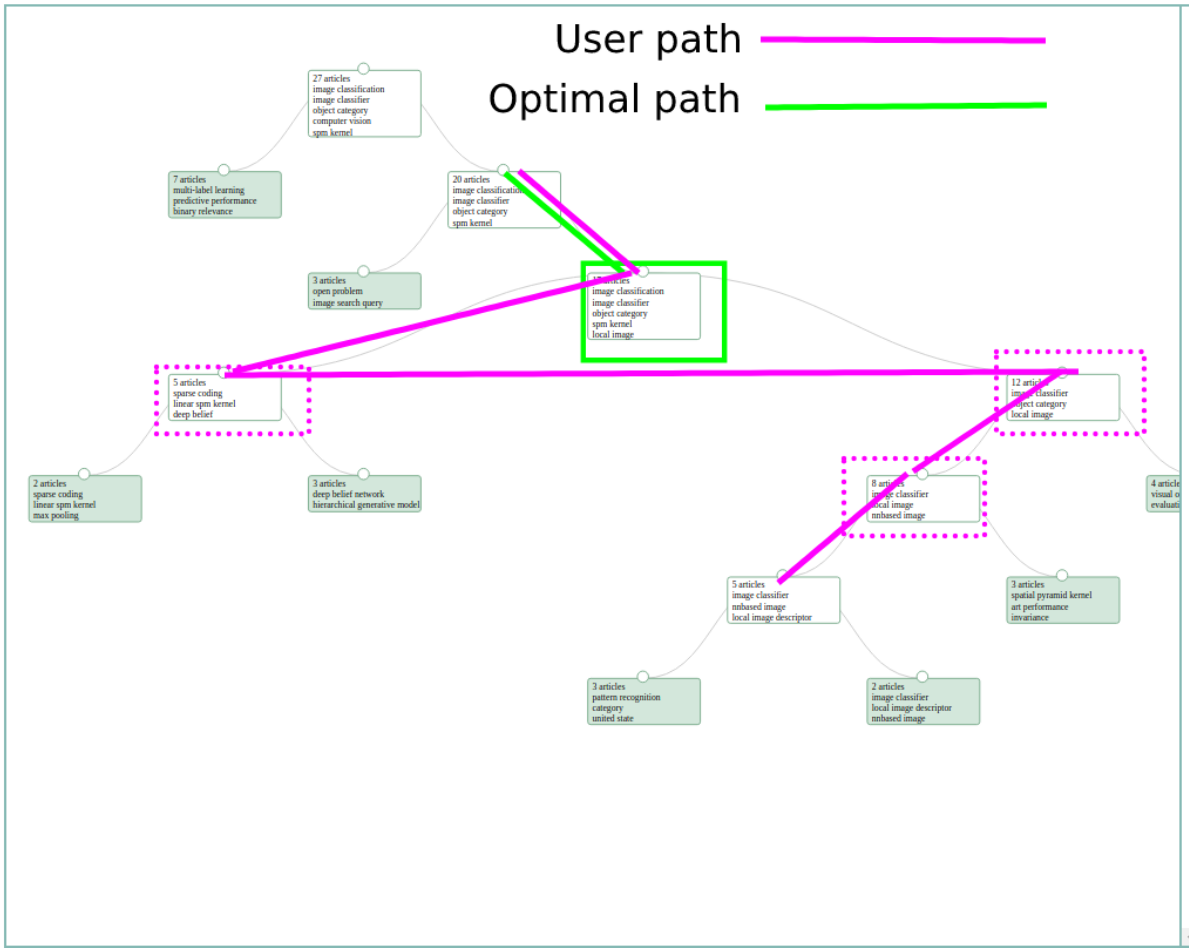
We compute the number of changing branches while expanding tree nodes.



Benchmarks: Attempts

How match attempts the user needs?

We compute number of user fails when choosing. If a user succeed on the first try, it equals to 0.



Locality-constrained Linear Coding for Image Classification

The traditional SPM approach based on bag-of-features (BoF) requires nonlinear classifiers to achiev

PDF

Authors: Wang Yang Yu Lv Huang

2010

Self-taught learning: Transfer learning from unlabeled data

We present a new machine learning framework called" self-taught learning" for using unlabeled data i

PDF

Authors: Raina Battle Lee Packer Ng

2007

The pyramid match kernel: Discriminative classification with sets of image features

Discriminative learning is challenging when examples are sets of features, and the sets vary in card

PDF

Authors: Grauman Darrell

2005

Compared algorithms

- HCBasic
- ComboBasic (Astrakhntsev, 2016)
- hierMTWL$_{idf}$ (Muhr, 2010)
- MTWL$_{idf}$ (Muhr, 2010)

Volunteers' datasets

Table 1: Datasets

Dataset number	**Properties**	
	Field	***Sampled size***
1	Graph data-bases	17
2	Web page data extraction	18
3	Social network graphs	20
4	Generating similar graphs	23
5	Cascades	29
6	Clustering	34
7	Exploratory search	56
8	Active learning	67

Results: Total statistics

Table 2: Total averages of benchmarks

Labelling algorithm	Answers amount	Benchmarks (average)		
		PathRatio	*Attempts*	*Jumps*
hierMTWL$_{idf}$	86	4.47	3.63	3.30
MTWL	48	3.00	4.17	3.13
ComboBasic	70	3.51	3.86	3.26
HCBasic	91	3.55	3.96	3.07

- The significance level of collected data were not high enough
- The numbers per algorithm were very different for different datasets

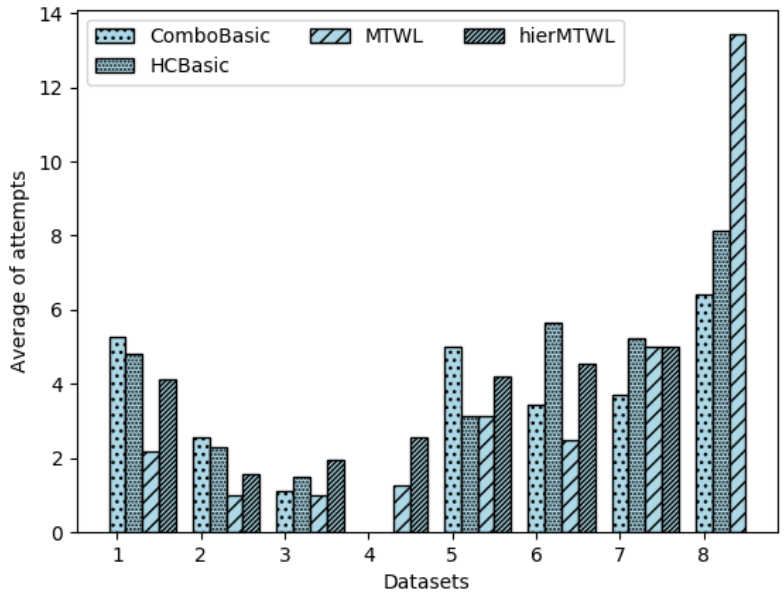


Figure 4: Average of attempts over the each collection

Partial findings: How many attempts do the users need?

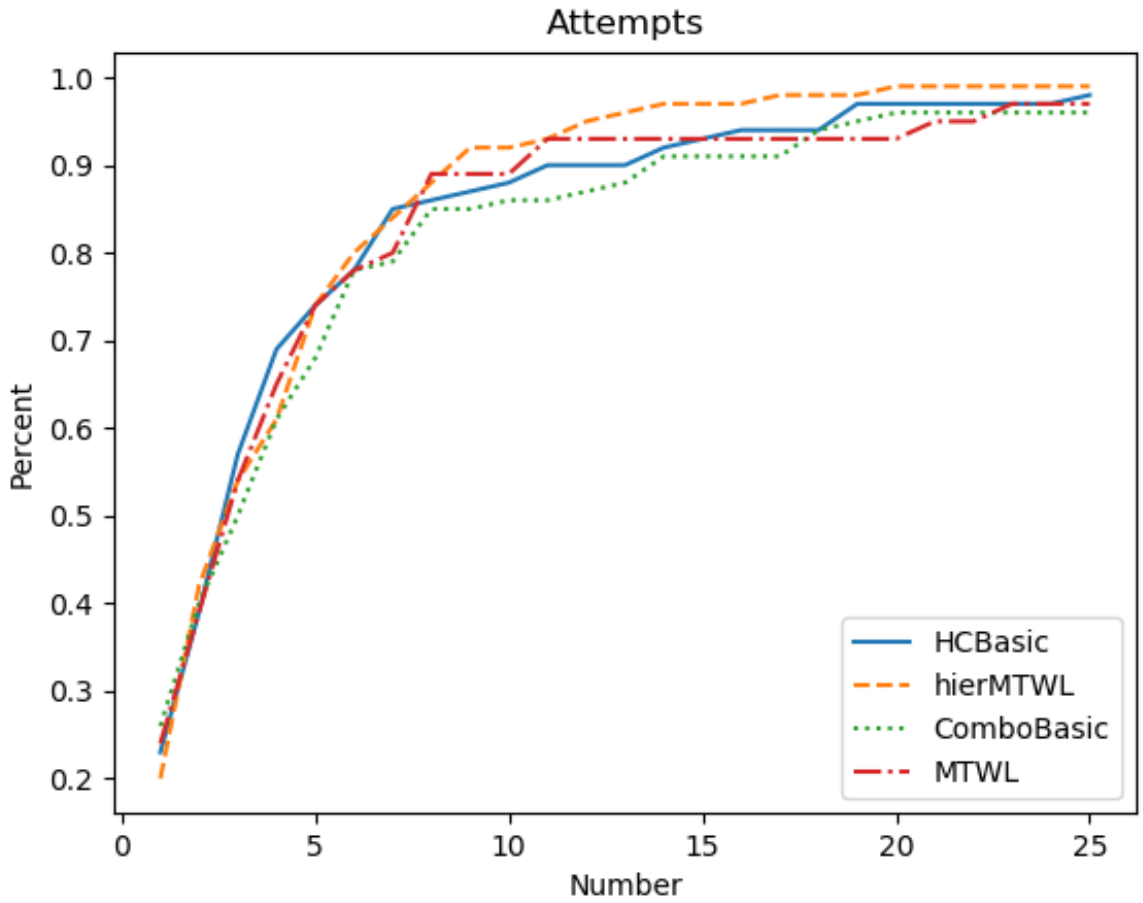


Figure 5: Required number of attempts before a correct answer

Partial findings: How frequently do the users take a wrong branch?

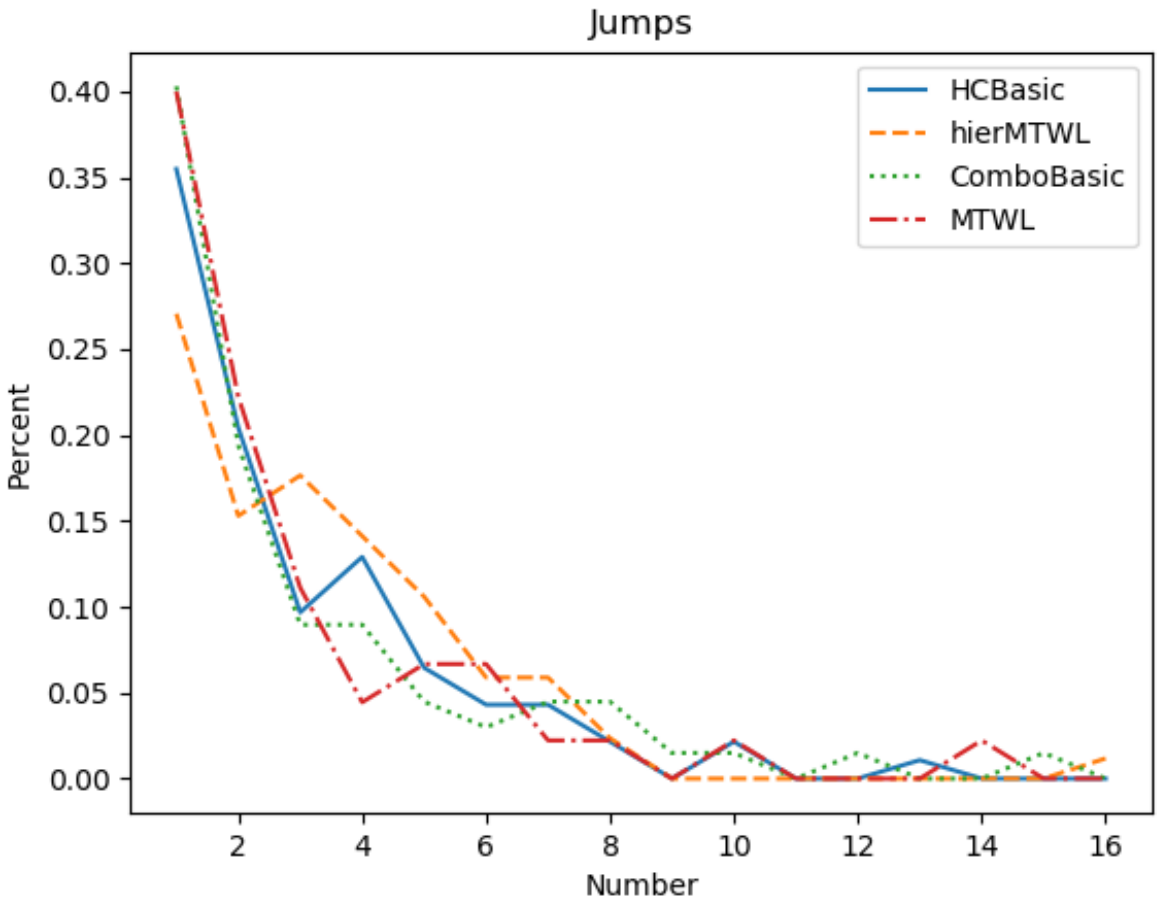


Figure 7: Number of jumps done by the participants before a correct answer

Conclusion[1]

1. HCBasic labelling method

- cluster position in hierarchy explicitly sets out the level of specificity of labels;
- designed especially for articles' abstracts

2. New evaluation strategy

- "in vivo"
- checking the requirements for clustering description implicitly

[1]The reported study was partially funded by RFBR according to the research project 17-07-00978 A.