A Combined Approach to **Dynamic web page Classification**: merging **Structure** and **Content**

Maria Niarou, Sofia Stamou

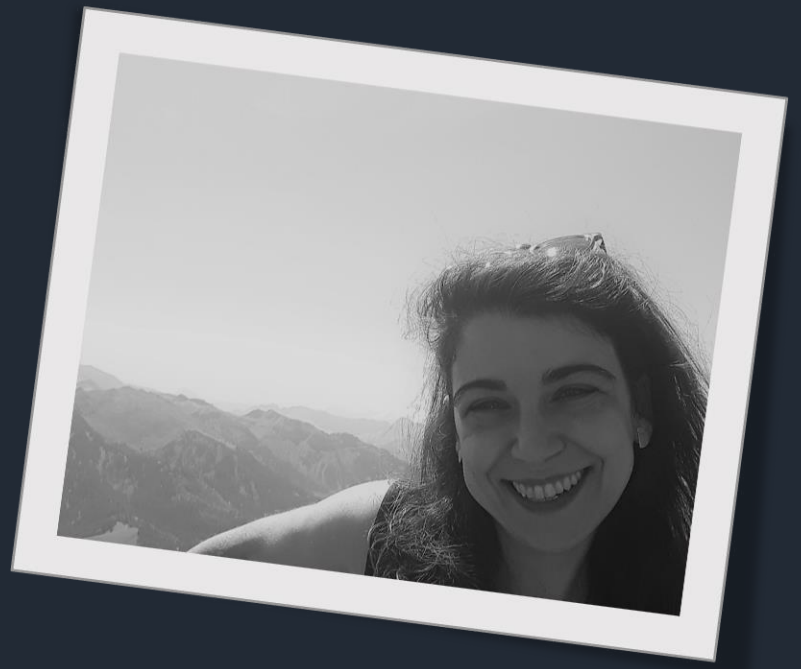
Department of Archives, Library and Museum Studies
Ionian University
Corfu, Greece

[l14niar@ionio.gr, stamou@ionio.gr]

WEB2020, IARIA Conference
Lisbon, Portugal

Presenter:

Maria (l14niar@ionio.com) is a PhD student at the Ionian University in Corfu. She is a graduate of the Department of Archives, Library Science and Museology (Faculty of Information Science & Informatics). During her Master's studies (Information Service in Digital Environment), she concluded that what interests her most is the data processing via semantic analysis. The research she conducted during the writing of her master's thesis resulted in her first publication: Niarou M., Stamou S. 2012; "Exploring Lexical Ontologies for Hierarchically Organizing the Greek Wikipedia Articles"; International Journal of Digital Information Management, Vol. 10, No. 3, pp. 157-167. The present presentation is the core of the academic research that Maria carries out, supervised by Sofia Stamou (Assistant Professor at the Department of Archives, Library Science and Museology).



Research Interests:

- Natural Language Processing
- Information Retrieval
- Semantics
- Knowledge Extraction from Digital Libraries and Web
- Knowledge Organization Systems

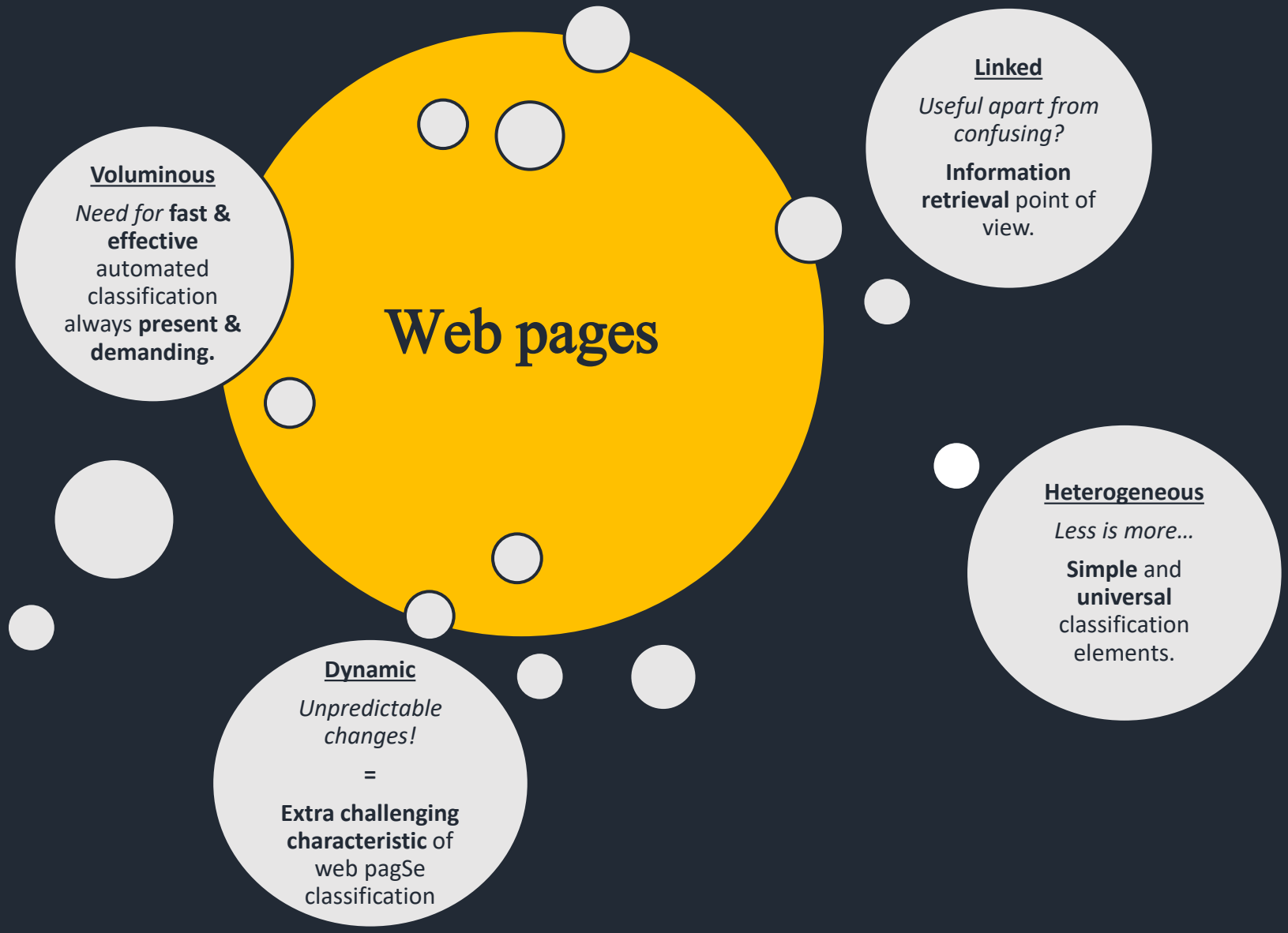
Languages: Greek, English, Italian, French

Presentation Layout

- INTRODUCTION
  - *Motivation*
  - *Related Work*
  - *Contribution*
- METHODOLOGY
  - *Outline*
  - *Algorithm1: Multi-Dimensional Classification (Procedure1)*
  - *Algorithm1: Multi-Dimensional Classification (Procedure2)*
  - *Algorithm2: Re-Classification based on Pages' Change Detection*
  - *Algorithm3: Optimized Re-Classification based on Change's Frequency Detection*
- EXPERIMENTAL STUDY
  - *Framework*
  - *Results*
- CONCLUSION and FUTURE WORK
  - *Conclusion*
  - *Future Work*

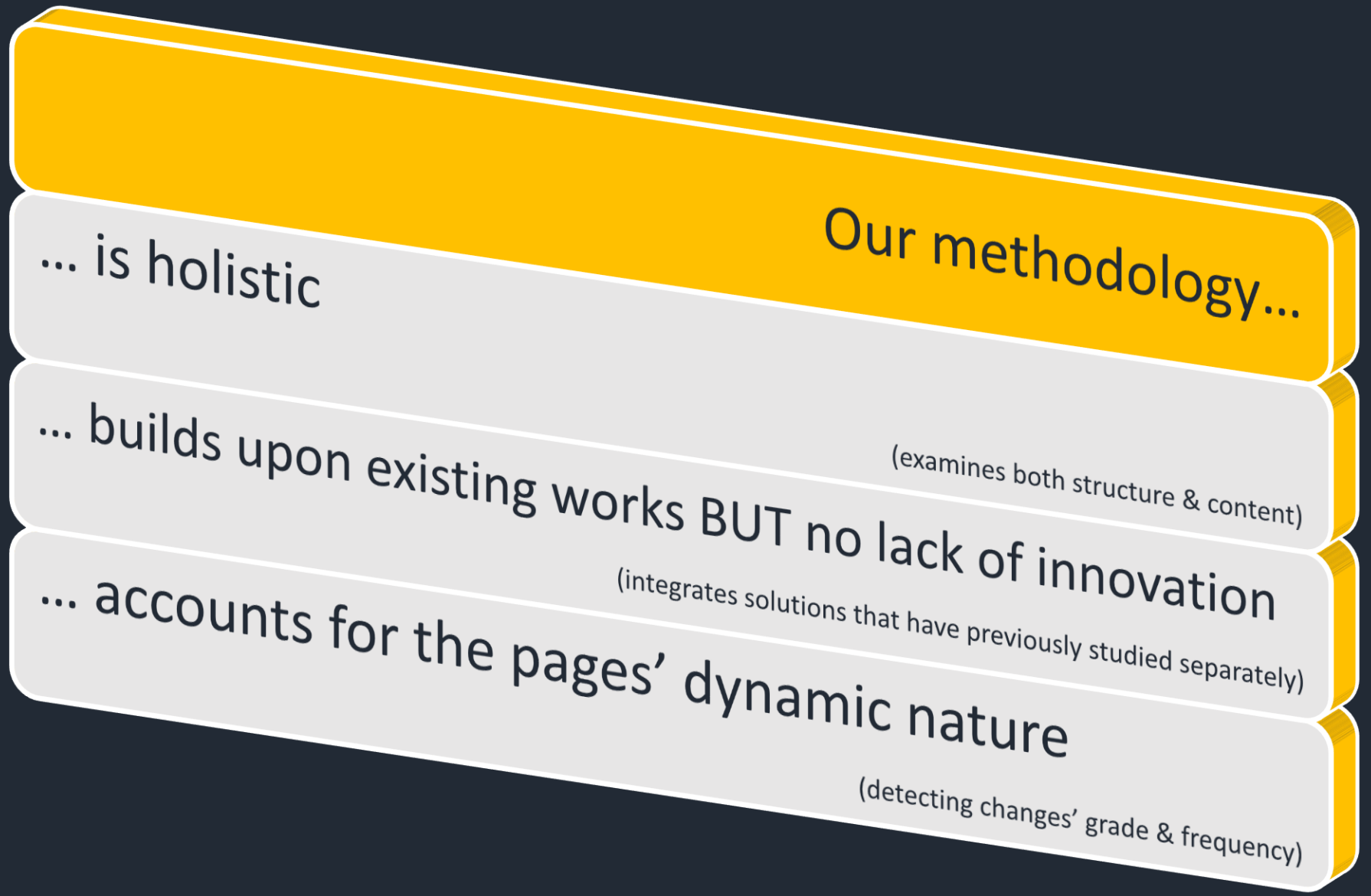
❖ Intro

Motivation: Web pages' nature (4 basic features)



Related Work (2 main groups)

Contribution



❖ Methodology

Outline : three autonomous and supplementary algorithms

Algorithm 1 : Multi-Dimensional Page Classification

- Idea: textual & structural classification = complementary
- Goal: to classify the pages structurally and thematically

Algorithm 2: Re-Classification based on Changes' Grade Detection

- Idea: detect pages' dynamic nature via a re-Classification component, in order to *maintain data indexes updated*
- Goal: to decide which of the modified pages need to be re-Classified

Algorithm 3: Optimized Re-Classification based on Changes' Frequency Detection

- Idea: optimize the re-Classification process
- Goal: save time and resources

Algorithm 1: Multi-Dimensional Page Classification

Procedure 1: Structure-based classification

Baseline: *we reverse the argument that web searches can be classified as either Navigational, Informational or Transactional, and we claim that pages can be classified accordingly.*

Input elements
"Clue" classification elements

- Phase1: Page Type Recognition

if any transactional term(s) as link(s) =
transactional page
Else
compute word tokens to links ratio (R)

if R ≥ t, P = P(informational)
Else
if R ≤ t , P = P(navigational)

Output →

Pages' Type identification
- P(transactional)
- P(navigational)
- P(informational)

- Phase2: Layered Page Classification given their type

Transactional Pages

- map P's(transactional) t(terms) to table correlation
- estimate the mapping(s) occurrence
- tag P(transactional) with the category of max occurrence (e.g. e-commerce)

Else
Search for payment term(s) as link(s)
if any, P(transactional) = "not-free"
Else
P(transactional) = "free"

Navigational Pages

starting after "http(s)://",
count the number of "/" in url
if "/" ≥ h, P(navigational) = "WebPage"
Else
P(navigational) = "HomePage"

HomePages
URLs mapping against (D(top))
and
tagged with their suffix meaning

WebPages
number of "/" = depth value

Informational Pages

*(no extra typical characteristics =>
no extra structural classification)*

Structurally classified pages

Algorithm1: Multi-Dimensional Page Classification

Procedure2: Content-based classification

Baseline: *having experimented with several textual features, we ended up with* ***anchor title*** *and* ***title*** *as the most easily extracted and informative of the theme of a page.*

Input elements
"Clue" classification elements

- **Phase1: Textual elements Extraction**

for each P,
extract anchor title in Url
And
title in text body

Output

Pages tagged by their textual elements

- **Phase2: Theme Detection**

for every page P,
-look for common terms between P's anchorTitle and P's textTitle
if found, use common terms as the theme(-s) to tag P
Else
-check for overlapping terms between P's first n-appearing keywords and (P's anchor title and P's text title)
if found, use overlapping terms as the theme(-s) to tag P
Else
-map P's first n-appearing keywords to WordNet and look for common senses between P's keywords and (P's anchor title and P's text title)
if found, use terms of common senses as the theme(-s) to tag P

Output

Thematically classified pages

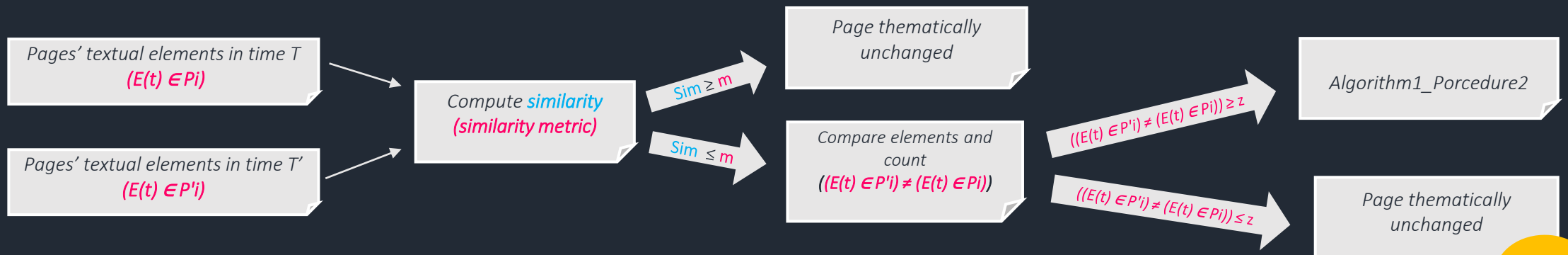
Structurally classified pages
+
Thematically classified pages
=
MultiDimensionally classified pages

Algorithm2: Re-Classification based on Change Detection

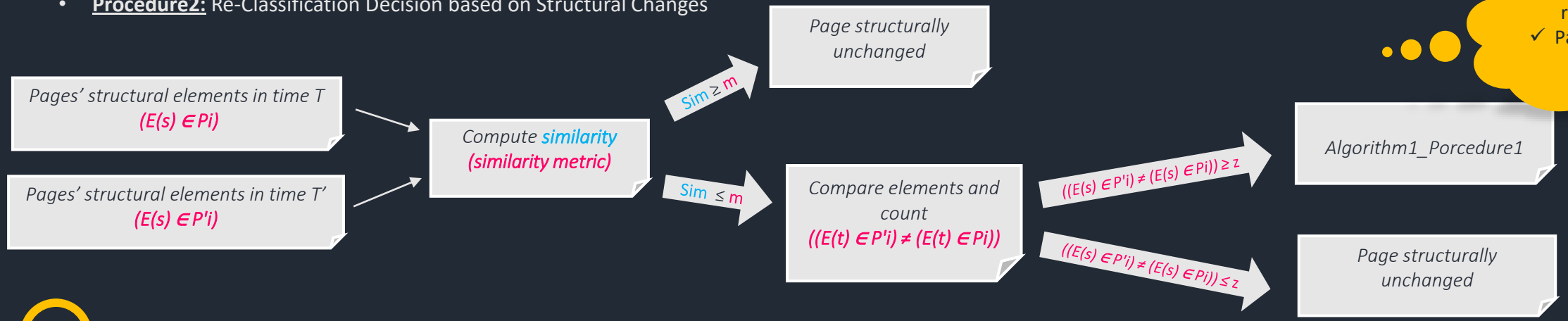
Challenge: how to deal with pages' dynamic nature // how to ensure that the classification outcome is up-to-date
Idea: re-Classification component to our algorithm
Goal: to detect, measure and identify the *possible* changes

Input elements
"Clue" classification elements

- **Procedure1:** Re-Classification Decision based on Textual Changes



- **Procedure2:** Re-Classification Decision based on Structural Changes



✓ Pages that need re-Classification
✓ Pages unchanged over time

Compare (structural and textual) elements of any given page with their counterparts previously identified!

Algorithm3: Optimized Re-Classification based on Change's Frequency Detection

Baseline:
capture pages' change frequency **=>**
=> determine the re-classification policy => optimize the runs of our Re-Classification algorithm =>
=> save time and resources

- activated upon the Re-Classification algorithm initializes
- records every time a change is detected (between two chronologically different snapshots of the page)

- changes are recorded by the timer
- timestamp of the change detection

MinFreqChange ≤ Ts ≤ MaxFreqChange

Send page to Algorithm2

Ts ≥ MaxFreqChange

Ts ≤ MinFreqChange

Secondary index

Secondary index

Input elements
"Clue" classification elements

Pages that need periodical Re-Classification

❖ Experimental Study

Framework (4 points)

Goal

- *Evaluate the effectiveness of our classification algorithm.*

How?

- Validating the classification performance of our method.
- Validate the potential weaknesses of our approach.

Data Set

- Bookmarked pages of 10 volunteers, experienced web users.
- Informed about our study objectives,
  asked to indicate pages' type & theme.
- - Brief instructions with respect to the definition of every type
  and
  - Several examples
- Theme self-determined by our volunteers
  and
  Verbalized based on their understanding

Quantity Formulas

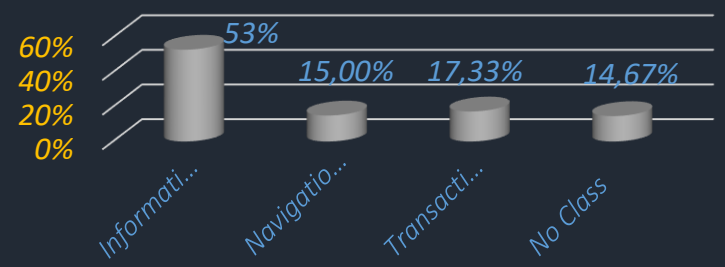
- Recall: $\frac{TP}{TP+FN}$
- Precision: $\frac{TP}{TP+FP}$

Experimental Study (2/2)

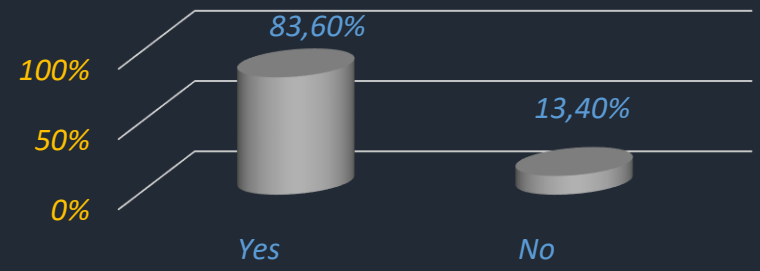
Results

Recall: algorithm's capacity in identifying a category for every page.
Precision: algorithm's capacity in identifying the correct category for every page.

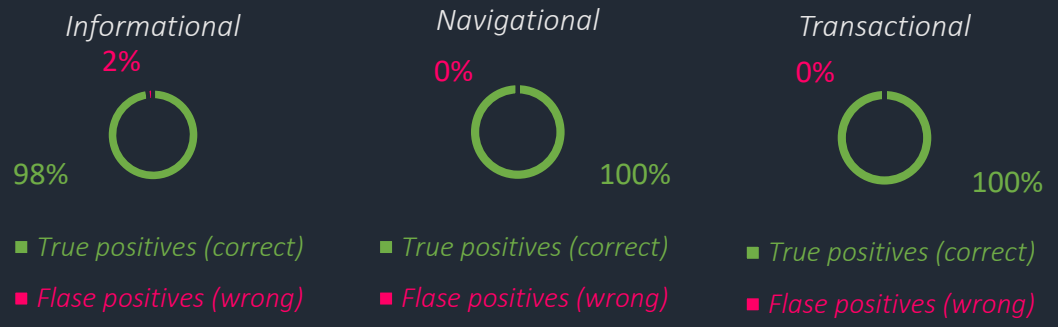
Structure-based Classification
(recall)



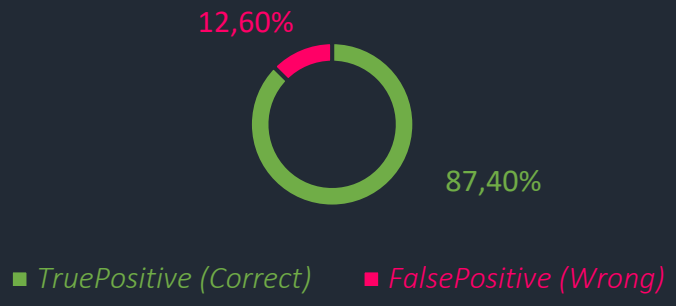
Content-based Classification
(recall)



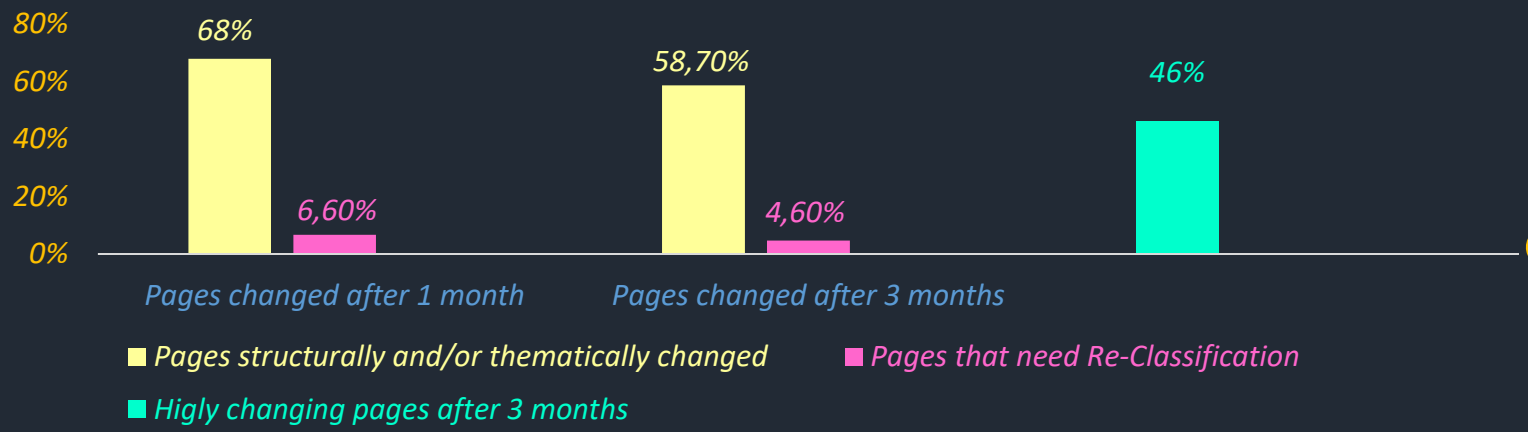
Structure-based Classification
(precision)



Evaluation of Content-based Classification
(precision)



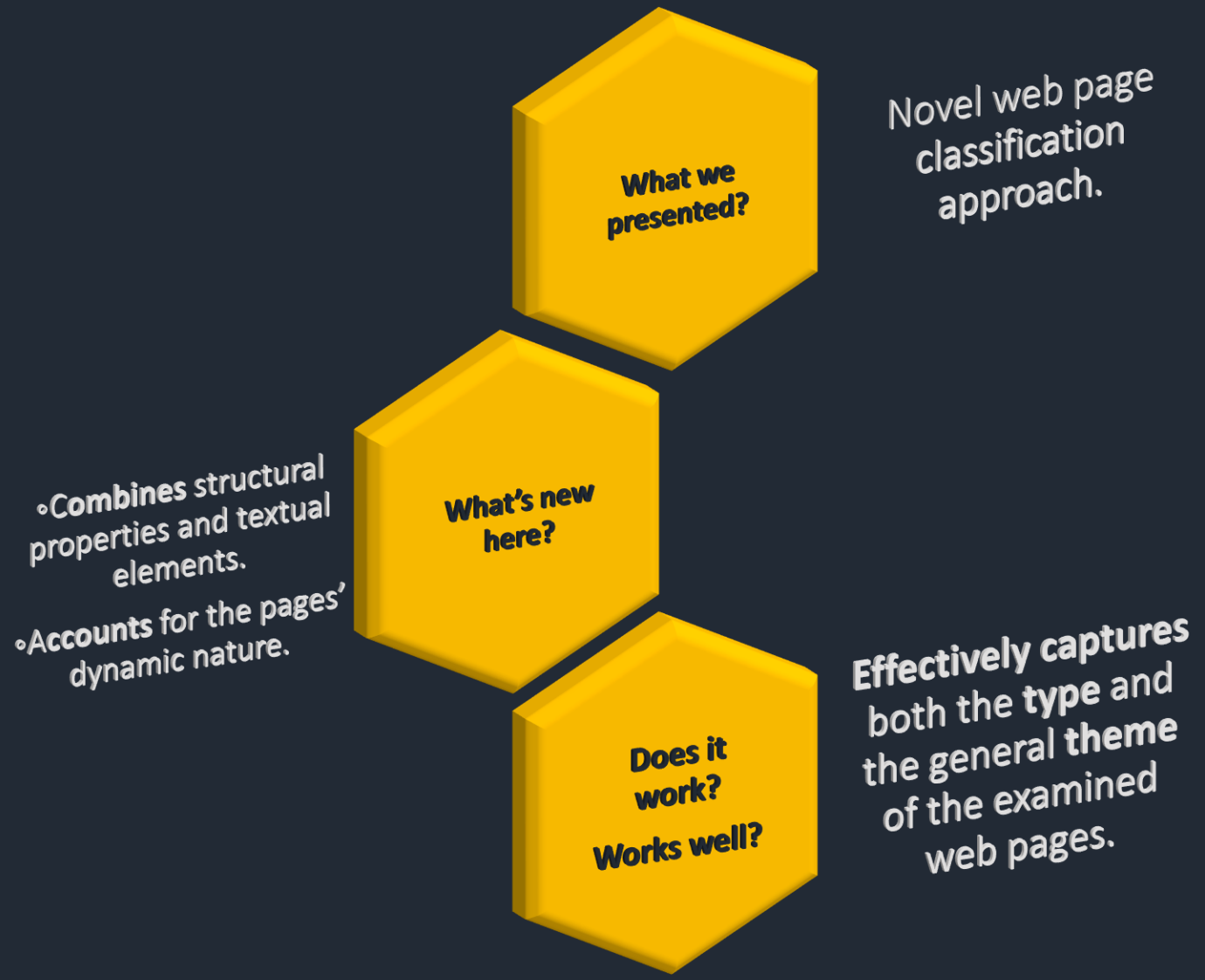
Changing pages that need Re-Classification



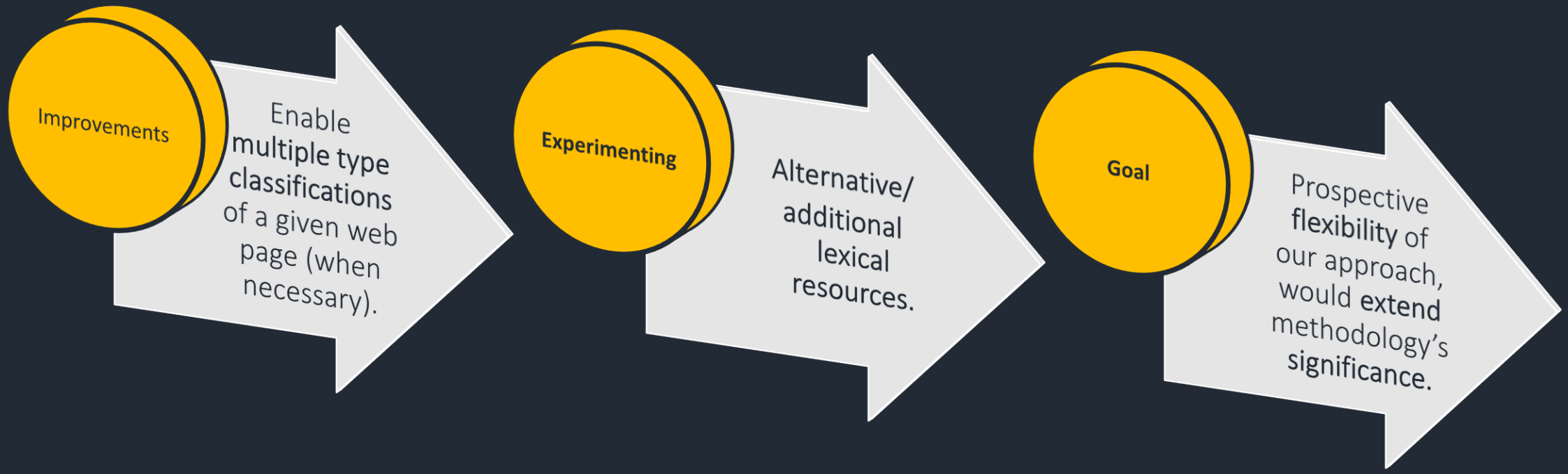
The algorithm successfully detects a *structural* & a *thematic* category for the majority of the examined pages!

❖ Conclusion & Future Work

Conclusion


What we presented?
Novel web page classification approach.
What's new here?
◦Combines structural properties and textual elements.
◦Accounts for the pages' dynamic nature.
Does it work?
Works well?
Effectively captures both the type and the general theme of the examined web pages.

Future work


Improvements
Enable multiple type classifications of a given web page (when necessary).
Experimenting
Alternative/ additional lexical resources.
Goal
Prospective flexibility of our approach, would extend methodology's significance.

Any Questions…

A Combined Approach to
Dynamic web page Classification:
merging **Structure** and **Content**

Maria Niarou, Sofia Stamou

Department of Archives, Library and Museum Studies
Ionian University
Corfu, Greece

[l14niar@ionio.gr, stamou@ionio.gr]

WEB2020, IARIA Conference
Lisbon, Portugal

Thank you!