



# IDENTIFYING SIMILAR LEARNING OBJECTS INCREMENTALLY

---

NAVEEN NELAPUDI

COMMITTEE:

Dr. SUSAN GAUCH (CHAIR)  
Dr. JERZY GRZYMALA BUSSE  
Dr. XUE-WEN CHEN



# OUTLINE OF THE PRESENTATION

---

- Introduction
- IKME
- Goals
- Overview
- Implementation
- Evaluation
- Screen Shots
- Conclusions
- Future Work



# Introduction

---

- According to Institute of Higher Education Policy, 85% of four year colleges offer online courses.
- Many institutions offer similar courses, so there are probably hundreds of descriptions of similar topics on the Internet.
- Educational content is expensive to produce.
- Courses are not very flexible and cannot be easily re-purposed.



# LEARNING OBJECTS

---

- Definition of Learning Object.
- Characteristics of a Learning Object.
  - Smaller Units of Learning.
  - Self Contained.
  - Reusable.
  - Can be aggregated.
  - Tagged with meta data.
- Reusable and very flexible when compared to a course as a learning unit.

# IKME

---

- Intelligent Knowledge Management Environment (*IKME*) is an ongoing project at the *University of Kansas* aimed at assisting the *Defense Information Technology Test bed (DITT)/University After Next (UAN)* by providing an advanced reach-back capability for commanders, staff, and other users who have time-critical needs.
- A knowledge management environment would facilitate the creation of extensible and reusable learning objects that would lead to faster delivery of content to knowledge users.
- XML is used as the data format for publishing. Knowledge creators use the environment to create learning objects which are stored as XML documents.



# Why XML?

---

- Advantages of XML
  - Structured - Capable of representing an object hierarchy.
  - Machine-readable and writeable.
  - Separates content from presentation.
- So, Learning Objects can be easily integrated into online courses.



# Goals

---

- To help the user find the related learning object content for the creation of lesson objects and manuals.
- To use a Memory based approach for indexing as opposed to File based approach used for the earlier version.
- To incorporate Incremental Indexing into the similarity search instead of batch processing which required the algorithm to be re-run every time a new document is added to the collection.



# Overview

---

- Problems with Earlier Version
  - File based Indexing.
  - Need to re-run the whole algorithm when a new document is added.
- My enhancements.
  - Memory based.
  - Incremental Indexing.



# Formula

---

**Formula:**

$$\text{Similarity}(d1, d2) = \sum_{i=1}^N wt_{id1} * wt_{id2}$$

where

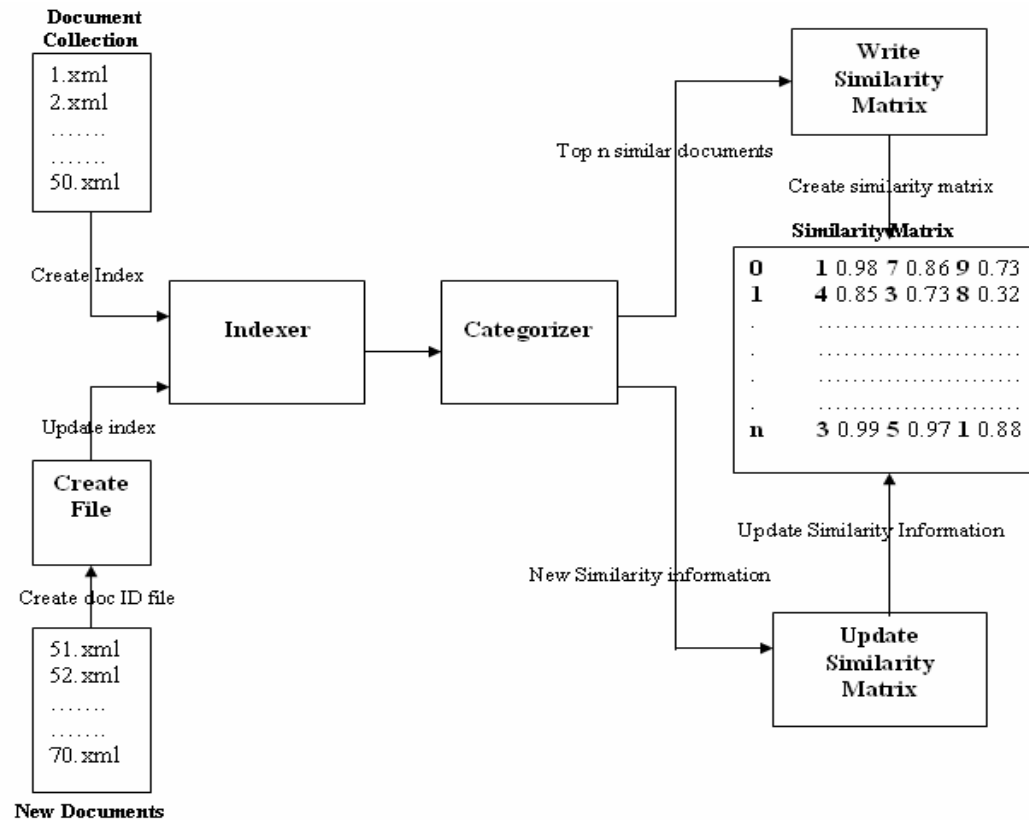
$N$  = number of tokens in the vocabulary

$$wt_{id1} = tf_{id1} * idf_i$$

$tf_{id1}$  = (frequency of token  $i$  in  $d1$  / total number of unique tokens in  $d1$ )

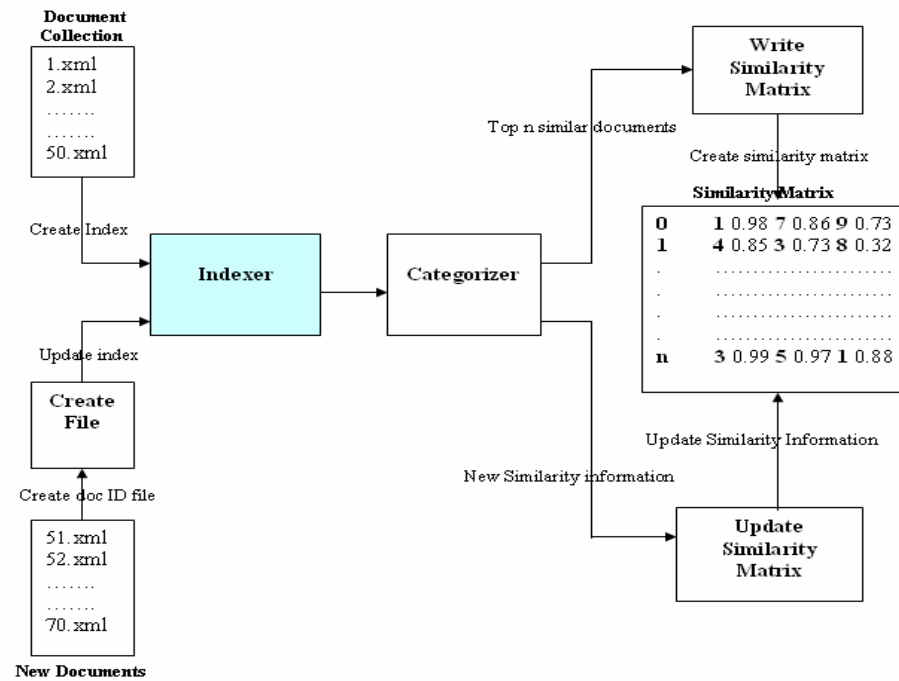
$idf_i$  =  $\log_2$  (total number of documents / number of documents in which token  $i$  appeared)

# System Architecture



System Architecture

# Indexer



System Architecture

Indexes the documents using standard vector space model (tf-idf) approach.



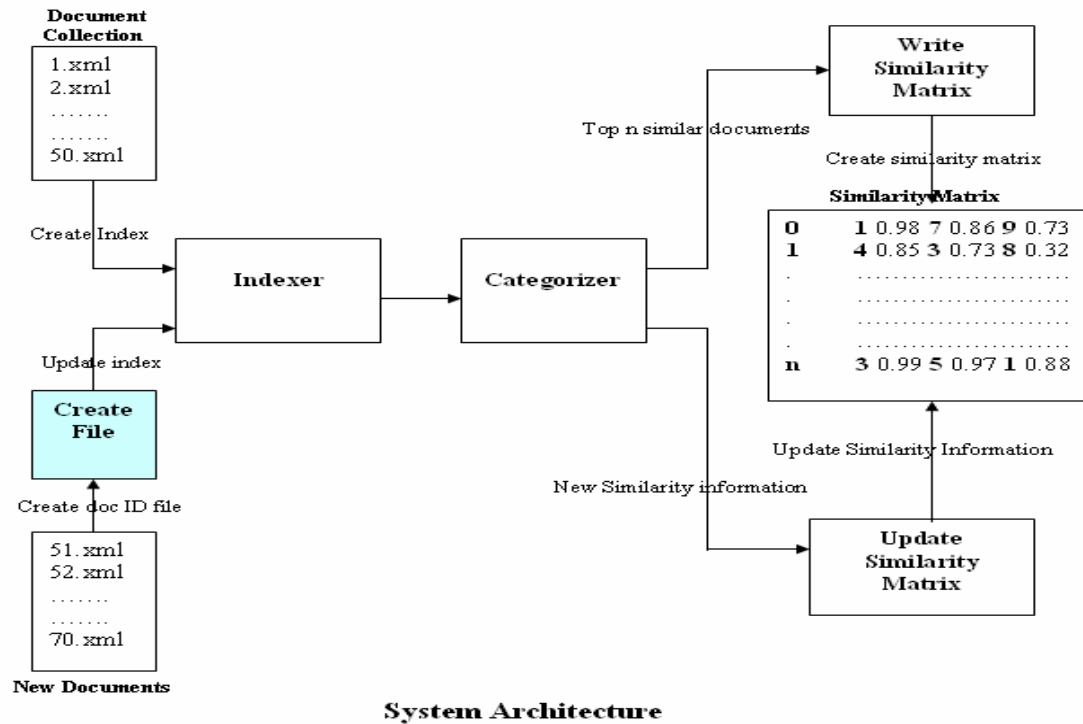
# Categorizer Output

---

## Example of Categorizer Output

Doc Id	Weight
10	0.981034
8	0.851032
6	0.752106
3	0.433210
11	0.121012
23	0.110012
1	0.090124
33	0.085291
42	0.067102
29	0.005128

# CreateFile



Creates a docID file which has document names and IDs. This file is needed for incremental indexing.

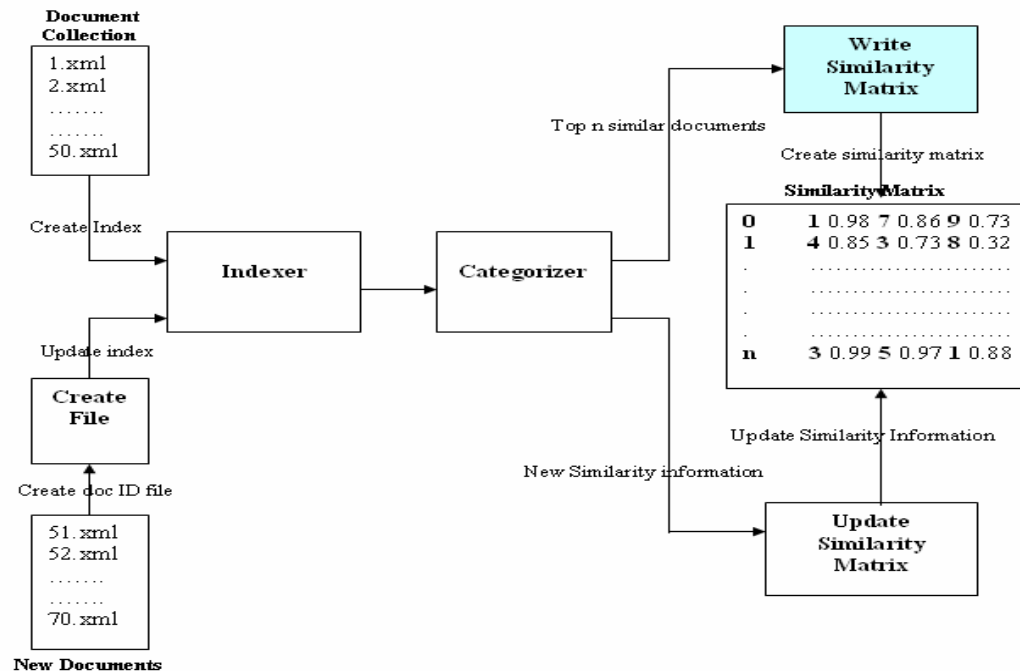
# Sample docID file

---

Example of docID file

Doc Name	Doc Id
../inputfiles/urban_energy.xml	51
../inputfiles/urban_environment.xml	52
../inputfiles/urban_finance.xml	53
../inputfiles/urban_infrastructure.xml	54
../inputfiles/urban_population.xml	55
../inputfiles/urban_network.xml	56
../inputfiles/urban_satellite.xml	57
../inputfiles/urban_segments.xml	58
../inputfiles/urban_supersurface.xml	59
../inputfiles/urban_systems.xml	60

# WriteSimilarity



System Architecture

Generates the Similarity Matrix.

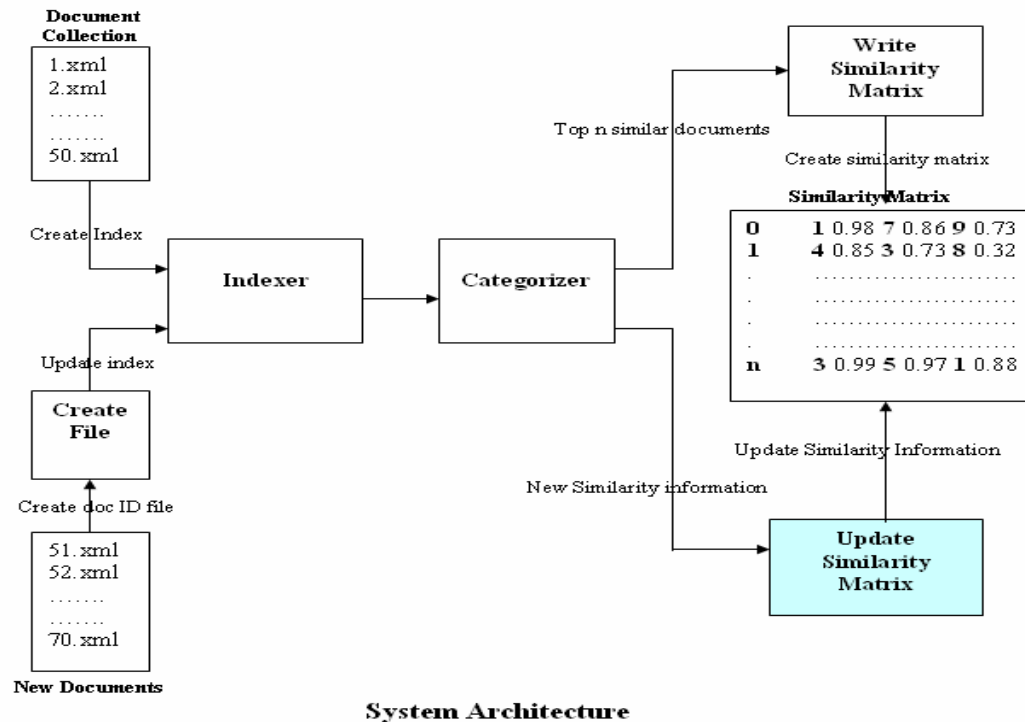


# Format of Similarity Matrix

---

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
1	31	0.231273	13	0.214116	46	0.078542	5	0.058139	7	0.052780
2	7	0.537214	13	0.448968	9	0.226499	46	0.223909	25	0.188427
3	49	0.128824	17	0.128532	13	0.092650	12	0.080430	50	0.075375
...	...	.....	...	.....	...	.....	...	.....	...	.....
...	...	.....	...	.....	...	.....	...	.....	...	.....
49	47	0.478480	7	0.405528	42	0.300528	42	0.269646	43	0.253155
50	12	0.248636	47	0.188142	46	0.180350	7	0.158472	48	0.134855

# Update Similarity



To update the similarity information, because of the newly added documents to the collection.

# STEP 1: Adding the new record(s) to the Similarity Matrix

---

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
<b>1</b>	31	0.231273	13	0.214116	46	0.078542	5	0.058139	7	0.052780
<b>2</b>	7	0.537214	13	0.448968	9	0.226499	46	0.223909	25	0.188427
<b>3</b>	49	0.128824	17	0.128532	13	0.092650	12	0.080430	50	0.075375
...	...	.....	...	.....	...	.....	...	.....	...	.....
...	...	.....	...	.....	...	.....	...	.....	...	.....
<b>49</b>	47	0.478480	7	0.405528	42	0.300528	42	0.269646	43	0.253155
<b>50</b>	12	0.248636	47	0.188142	46	0.180350	7	0.158472	48	0.134855
<b>51</b>	1	0.422716	49	0.412513	31	0.331092	42	0.290143	28	0.271103
<b>52</b>	2	0.610218	33	0.531274	41	0.442091	27	0.301208	8	0.200172
<b>53</b>	36	0.312082	3	0.229142	24	0.220081	46	0.193842	5	0.164032

# STEP 2: Updating the corresponding records' similarity information

51	1	0.422716	49	0.412513	53	0.331092	42	0.290143	28	0.271103
----	---	----------	----	----------	----	----------	----	----------	----	----------

Similarity (1,51) = 0.422716

Similarity (49,51) = 0.412513

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
1	31	0.231273	13	0.214116	46	0.078542	5	0.058139	7	0.052780
2	7	0.537214	13	0.448968	9	0.226499	46	0.223909	25	0.188427
3	49	0.128824	17	0.128532	13	0.092650	12	0.080430	50	0.075375
...	...	.....	...	.....	...	.....	...	.....	...	.....
...	...	.....	...	.....	...	.....	...	.....	...	.....
49	47	0.478480	7	0.405528	42	0.300528	42	0.269646	43	0.253155
50	12	0.248636	47	0.188142	46	0.180350	7	0.158472	48	0.134855
51	1	0.422716	49	0.412513	31	0.331092	42	0.290143	28	0.271103
52	2	0.610218	33	0.531274	41	0.442091	27	0.301208	8	0.200172
53	36	0.312082	3	0.229142	24	0.220081	46	0.193842	5	0.164032

# Updating Document 1's Similarity Information

---

## OLD RECORD OF DOC 1

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
1	31	0.231273	13	0.214116	46	0.078542	5	0.058139	7	0.052780

With the addition of new document 51 with Similarity  $(1,51) = 0.422716$

## NEW RECORD OF DOC 1

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
1	51	0.422716	31	0.231273	13	0.214116	46	0.078542	5	0.058139

# Updating Document 49's Similarity Information

---

## OLD RECORD OF DOC 49

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
49	47	0.478480	7	0.405528	42	0.300528	42	0.269646	43	0.253155

With the addition of new document 51 with Similarity  $(49,51) = 0.412513$

## NEW RECORD OF DOC 49

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
49	47	0.478480	51	0.412513	7	0.405528	42	0.300528	42	0.269646

# Updating the corresponding records' similarity information (contd')

52	2	0.610218	33	0.531274	41	0.442091	27	0.301208	8	0.200172
----	---	----------	----	----------	----	----------	----	----------	---	----------

Similarity (2,52) = 0.610218



Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
1	31	0.231273	13	0.214116	46	0.078542	5	0.058139	7	0.052780
2	7	0.537214	13	0.448968	9	0.226499	46	0.223909	25	0.188427
3	49	0.128824	17	0.128532	13	0.092650	12	0.080430	50	0.075375
...	...	.....	...	.....	...	.....	...	.....	...	.....
...	...	.....	...	.....	...	.....	...	.....	...	.....
49	47	0.478480	7	0.405528	42	0.300528	42	0.269646	43	0.253155
50	12	0.248636	47	0.188142	46	0.180350	7	0.158472	48	0.134855
51	1	0.422716	49	0.412513	31	0.331092	42	0.290143	28	0.271103
52	2	0.610218	33	0.531274	41	0.442091	27	0.301208	8	0.200172
53	36	0.312082	3	0.229142	24	0.220081	46	0.193842	5	0.164032

# New Similarity Matrix

---

Doc Id	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight	Doc Id	Weight
1	51	0.422716	31	0.231273	13	0.214116	46	0.078542	5	0.058139
2	52	0.610218	7	0.537214	13	0.448968	9	0.226499	46	0.223909
3	53	0.229142	49	0.128824	17	0.128532	13	0.092650	12	0.080430
...	...	.....	...	.....	...	.....	...	.....	...	.....
...	...	.....	...	.....	...	.....	...	.....	...	.....
49	47	0.478480	51	0.412513	7	0.405528	42	0.300528	42	0.269646
50	12	0.248636	47	0.188142	46	0.180350	7	0.158472	48	0.134855
51	1	0.422716	49	0.412513	31	0.331092	42	0.290143	28	0.271103
52	2	0.610218	33	0.531274	41	0.442091	27	0.301208	8	0.200172
53	36	0.312082	3	0.229142	24	0.220081	46	0.193842	5	0.164032



# Evaluation

---

- Incremental Updation of Similarity Information:

<b>Number of Documents</b>	<b>New Version</b> Time (in seconds)	<b>Earlier Version</b> Time (in seconds)
54	26 (7.31u, 12.28s)	52
72 (18 documents added)	11 (2.48u, 3.98s)	62
172 (100 documents added)	52 (14.13u, 23.10s)	220

For 500 documents (450 added to an initial collection of 50), it took around 243 seconds (68.34u, 124.36s) for the new version.

# Evaluation (cont'd)

---

- Calculating Similarity Information from the scratch:

<b>Number of Documents</b>	<b>New Version</b> Time (in seconds)	<b>Earlier Version</b> Time (in seconds)
54	26 (6.98u, 12.09s)	52
72	37 (9.418u, 15.85 s)	62
172	91 (24.21u, 38.09s)	220

# Screenshots – List of Learning Objects

The screenshot shows a Microsoft Internet Explorer browser window titled "The IKME Project - Microsoft Internet Explorer". The address bar contains the URL "http://onlineacademy.org/ikme/demoV1.0/demo\_menu.html". The website header features logos for the U.S. Army and KU (University of Kansas), along with the text "eLearning Design Lab Dole Center University of Kansas" and "IKME".

The main content area is organized into several sections:

- HOME**
- Learning Objects**
  - [Create](#)
  - [View\(with tags\)](#)
  - [View\(content only\)](#)
  - [Modify](#)
  - [Search](#)
  - [Import](#)
  - [Extended Search](#)
- Lesson Objects**
  - [Create](#)
  - [View\(with tags\)](#)
  - [View\(content only\)](#)
  - [Modify](#)
- Manual**
  - [Create](#)
  - [View](#)
  - [Modify](#)

The main content area lists the following learning objects:

- [Administration and Human Services](#)
- [An Urban Model](#)
- [Broad Urban Patterns](#)
- [Characteristics of Urban Operations](#)
- [Characteristics of Urban Operations](#)
- [Civilian Concerns](#)
- [Communications](#)
- [Comparison of Military Operations in Different Environments](#)
- [Definition of Joint Urban Operations](#)
- [Dimensions of Urban Terrain](#)
- [Energy Systems](#)
- [Financial](#)
- [General Population Size](#)
- [Group Size, Location and Composition](#)
- [Historical Overview of Military Urban Operations](#)
- [Impact on Future Operations](#)
- [Importance of Urban Areas in Future Military Operations](#)
- [Infrastructure System Interdependence](#)
- [Interior and External Spaces](#)
- [Linear Pattern](#)
- [Linear Urban Pattern](#)
- [Multidimensional Nature of Urban Terrain](#)
- [Network Pattern](#)
- [Network Pattern](#)

The browser's taskbar at the bottom shows the Internet Explorer icon and the text "Internet".

# Displaying the top 5 similar learning objects at the bottom

The screenshot shows a Microsoft Internet Explorer browser window displaying a web page titled "The IKME Project - Microsoft Internet Explorer". The address bar shows the URL: [http://onlineacademy.org/ikme/demoV1.0/demo\\_menu.html](http://onlineacademy.org/ikme/demoV1.0/demo_menu.html). The page features a green header with logos for the U.S. Army and KU, and text for "eLearning Design Lab Dole Center University of Kansas" and "IKME".

The main content area is titled "VIEW LEARNING OBJECT" and "Content". It contains a paragraph of text:

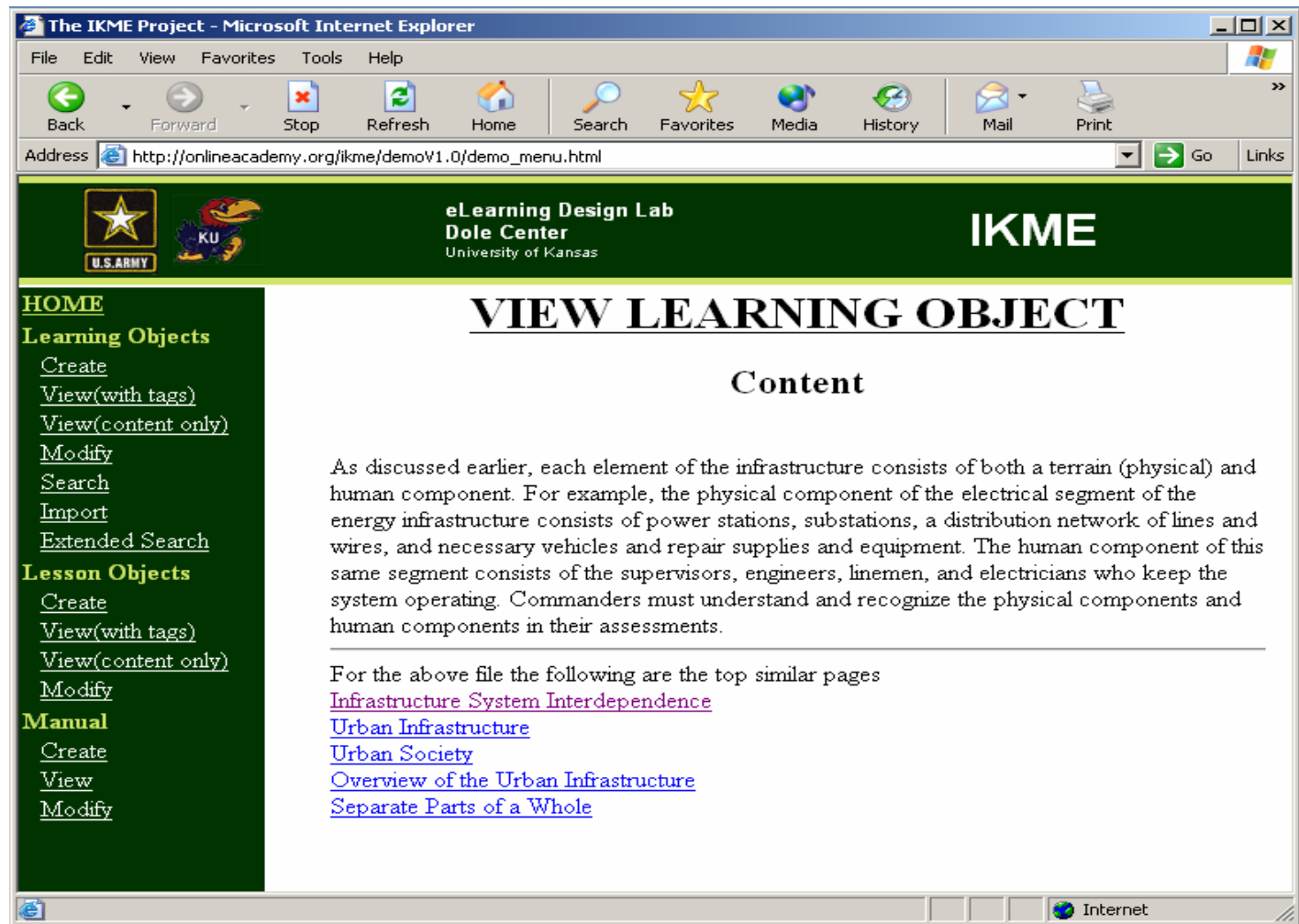
However, commanders must understand that destroying or disrupting any portion of the urban infrastructure can have a cascading effect (either intentional or unintentional) on the other elements of the infrastructure. Yet, they may be able to gain an operational advantage while minimizing unwanted effects using precision munitions, electronic disruption of communications, or SOF and conventional ground forces to seize or secure an essential facility or structure. To gain this advantage, commanders will rely more on the expertise of engineer and CA units, local urban engineers and planners, and others with infrastructure-specific expertise. After understanding the technical aspects of the area's systems, they can develop the most appropriate course of action.

For the above file the following are the top similar pages

- [Toxic Industrial Chemicals](#)
- [Financial](#)
- [Urban Dimensions](#)
- [Structures and People](#)
- [Urban Terrain](#)

The left sidebar contains navigation links for "HOME", "Learning Objects", and "Lesson Objects".

# Screenshots – Displaying the top similar learning object



The screenshot shows a Microsoft Internet Explorer browser window titled "The IKME Project - Microsoft Internet Explorer". The address bar displays the URL "http://onlineacademy.org/ikme/demoV1.0/demo\_menu.html". The browser's toolbar includes buttons for Back, Forward, Stop, Refresh, Home, Search, Favorites, Media, History, Mail, and Print. The website's header features logos for the U.S. Army, KU, eLearning Design Lab, Dole Center, and University of Kansas, along with the acronym "IKME".

The main content area is titled "VIEW LEARNING OBJECT" and "Content". It contains a paragraph of text:

As discussed earlier, each element of the infrastructure consists of both a terrain (physical) and human component. For example, the physical component of the electrical segment of the energy infrastructure consists of power stations, substations, a distribution network of lines and wires, and necessary vehicles and repair supplies and equipment. The human component of this same segment consists of the supervisors, engineers, linemen, and electricians who keep the system operating. Commanders must understand and recognize the physical components and human components in their assessments.

For the above file the following are the top similar pages

- [Infrastructure System Interdependence](#)
- [Urban Infrastructure](#)
- [Urban Society](#)
- [Overview of the Urban Infrastructure](#)
- [Separate Parts of a Whole](#)

The left sidebar contains navigation links under the heading "HOME":

- Learning Objects**
  - [Create](#)
  - [View\(with tags\)](#)
  - [View\(content only\)](#)
  - [Modify](#)
  - [Search](#)
  - [Import](#)
  - [Extended Search](#)
- Lesson Objects**
  - [Create](#)
  - [View\(with tags\)](#)
  - [View\(content only\)](#)
  - [Modify](#)
- Manual**
  - [Create](#)
  - [View](#)
  - [Modify](#)



# Conclusions

---

- The primary goal of incorporating Incremental Indexing into the similarity search has been achieved.
- The algorithm need not be re-run even if a single document is added to the collection. Only the required parts of the index and the similarity matrix are updated.
- This will provide a faster way to search for similar learning objects and help the educators in creating new lessons using existing learning objects rapidly and inexpensively.



# Future Work

---

- Investigating similarity formula (i e., weighting different fields differently when calculating the match between the objects)
- Learning best differential weighting scheme.

# References

---

- All about Learning Objects  
<http://www.eduworks.com/LOTT/tutorial/learningobjects.html>
- Learning Objects 101 : A Primer for Neophytes  
<http://online.bcit.ca/sidebars/02november/inside-out-1.htm>
- Introducing Reusable Learning Objects  
[http://media.wiley.com/product\\_data/excerpt/56/07879649/0787964956.pdf](http://media.wiley.com/product_data/excerpt/56/07879649/0787964956.pdf)
- “Automatically Identifying Related Learning Objects”  
Mahesh Vulpala, Masters Project. University of Kansas  
2003.





# Questions

---

- ???