Microsoft Research Silicon Valley

Mining Videos from the Web for Electronic Textbooks

R. Agrawal, M. Christoforaki, S. Gollapudi,
A. Kannan, K. Kenthapadi, & A. Swaminathan

Microsoft Research – Silicon Valley

June 11, 2014

Microsoft Research

ICFCA 2014
12th International Conference on Formal Concept Analysis

Outline

- Motivation
- Approach
- Experiments
- Concluding Remarks

Outline

- Motivation
- Approach
- Experiments
- Concluding Remarks

Data Mining for Enriching Electronic Textbooks

Diagnostic tools for identifying weaknesses in textbooks

Within section deficiencies

Syntactic complexity of writing and dispersion of key concepts in the section [AGK+11a]

Across sections deficiencies

Comprehension burden due to non-sequential presentation of concepts [ACG+12]

Algorithmic enhancement of textbooks for enriching reading experience

References to selective web content

Links to authoritative articles [AGK+10], images [AGK+11b] and videos [ACG+14] based on the focus of the section

References to prerequisites

Links to concepts necessary for understanding the present section, derived using a model of a how students read textbooks [AGK+13]

- Validation on textbooks from U.S.A and India, on different subjects, across grades
- Prototypes and research papers (see **References**)

Microsoft Research

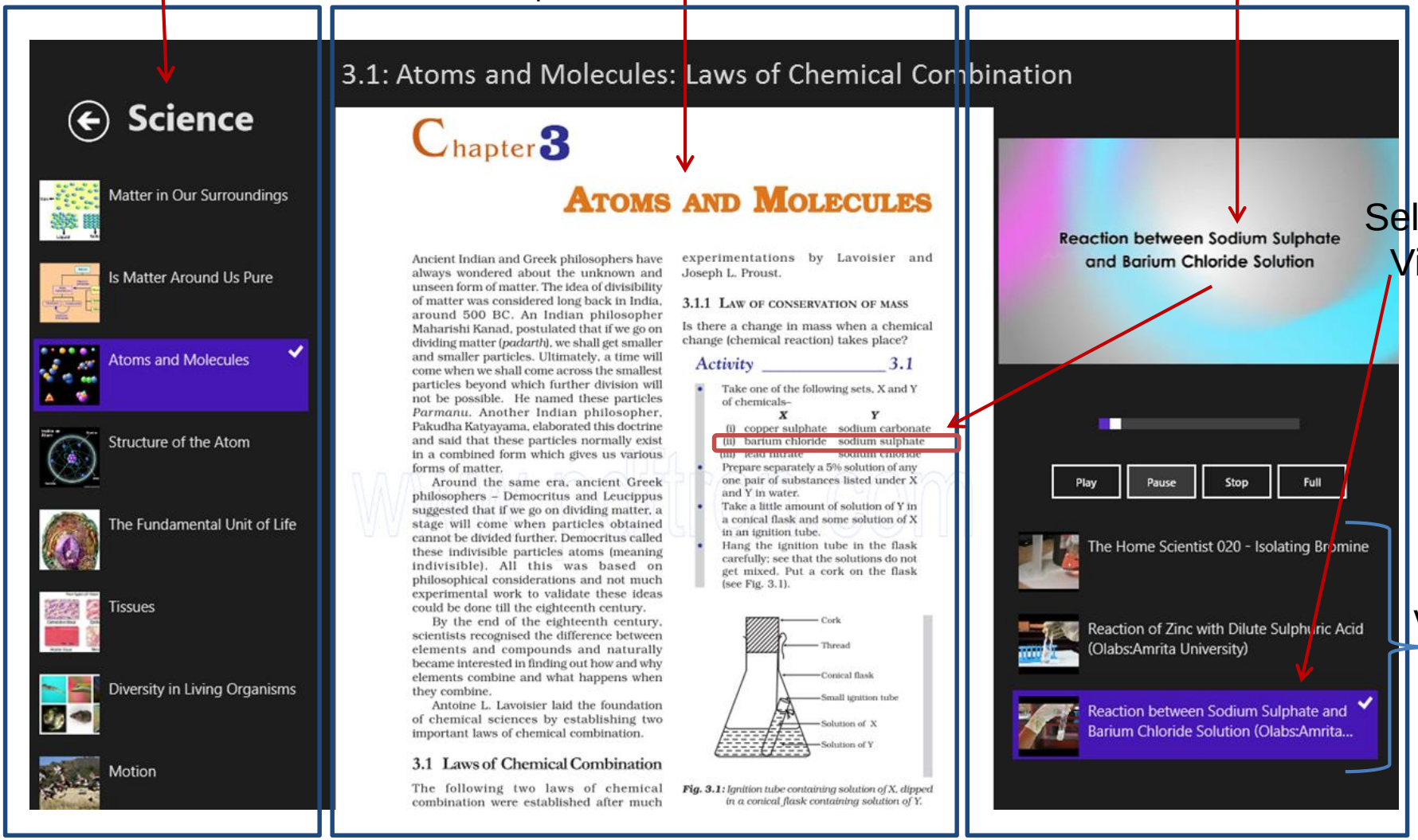
Joint work with Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, *et al.*

Video Augmentation: Make inaccessible accessible

Table of contents for navigating the book *(automatically extracted)*

Re-rendered section: This section, about the laws of chemical combination, prescribes an activity for the chemistry lab, but the school might lack the lab to do the experiments

Augmentations panel: Video demonstrates the reaction for the second set of chemicals prescribed



Selected Video

videos

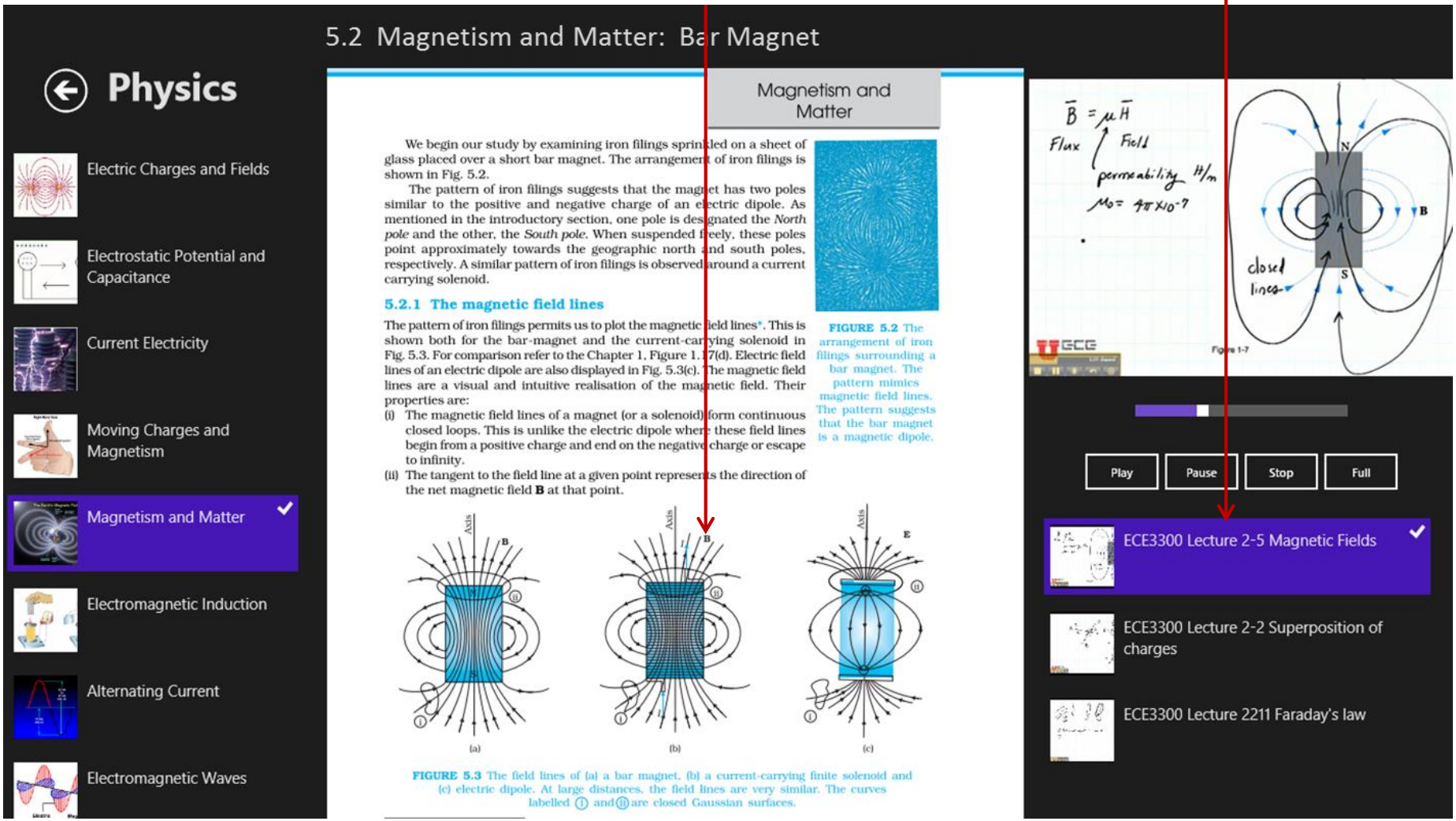
Microsoft Research

Win8 Surface Prototype

Video Augmentation: Assist in understanding content

This section is about magnetic field lines created by bar magnet. Section contains static images of magnetic field for bar magnet, solenoid and dipole.

The videos describes step-by-step magnetic field creation in bar magnet.



Microsoft Research

Win8 Surface Prototype

Outline

- Motivation
- Approach
- Experiments
- Concluding Remarks

Well-Written Textbooks

Two organizational principles [GL35, CC98]

[1] Focus:

- Each section explains a few concepts

[2] Unity:

- For each concept, there is a unique section that explains the concept

Focus = Unique contribution of the section

Methodology

- Compute focus of every section of the textbook
- Crawl the Web for educational videos [CBD99, S08]
- Obtain transcripts of the videos [SCF+06]
- Score transcripts with respect to the focus of the section

Methodology

- Compute focus of every section of the textbook
- Crawl the Web for educational videos [CBD99, S08]
- Obtain transcripts of the videos [SCF+06]
- Score transcripts with respect to the focus of the section

Seeking Focus

- Formal concept: Unit of human thought [W05]
- Focus of section *s*= {Formal concept} in *s*!
- Formal concepts being abstract, they manifest in the form of concept phrases, *cphrs*
- Extent = {Paragraph}, Intent = {*cphr*} in *s*
- A formal concept of s: A pair of maximal set of paragraphs Pc and maximal set of cphrs C s.t. every cphr in C is present in every paragraph in Pc
- FC1 = <{P1, P2, P3}, {x, y}>, FC2 = <{P3}, {x, y, z}>. Now what?

Refinement

- Iceberg lattice [S+02]: Robust to small perturbations
- A concept is frequent if at least μ objects individually have every attribute in the intent of the context
- Iceberg concept lattice = partial order over frequent concepts
- Use the leaf nodes of the iceberg concept lattice to represent focus (most specific sets of *cphrs*, or equivalently maximal combinations of *cphrs*)!

Focus of a Section

cphr: A concept phrase present in the text

Indicium of a section *s*: Set of *cphrs C* ∈ *s* s.t.

- *C* is frequent in *s*
- No section *s'* in which *C* is frequent
- No superset of *C* is frequent in *s*

Focus of a section s: Set of indicia present in *s*

Indicium & Class Hypothesis

- Hypothesis for a class [K04]: A closed itemset occurring in a class but not in other classes
- Minimal hypothesis: Inclusion minimal hypothesis
- Indicium = Maximal frequent (and hence closed) itemset in a class (text section), but infrequent in other classes (sections)
- Indicium = Relaxation of a minimal hypothesis, admitting infrequent presence in other classes

Indicia from Two Physics Sections

(a) Magnetism & Gauss' Laws

field line, magnetic field, monopole
field line, magnet, charged particle
electrostatics, field line, monopole

Section on general magnetism discussing magnetic fields and their effect on moving particles

'Monopole': hypothetical particle that is an isolated magnet with one pole; distinguishes this section on general magnetism from section on earth' magnetism

(b) Earth's Magnetism

field line, magnetic field, earth
equator, meridian, southern hemisphere
earth, solar wind, poles

Discusses how earth acts as magnet

Earth has both poles

'Earth' is generic term but indicium formed by combining it with 'field line' and 'magnetic field' very pertinent for the section

Microsoft Research

See paper for examples from the Biology book

Methodology

- Compute focus of every section of the textbook
- Crawl the Web for educational videos [CBD99, S08]
- Obtain transcripts of the videos [SCF+06]
- Score transcripts with respect to the focus of the section

Indicia of Transcripts?

- Videos produced independently and with different viewpoints
- Lack focus and unity properties (Not surprising)

Scoring Transcripts

- Candidate Video V = Transcript of V contains all cphrs from at least one indicium
- Score(V) = $\sum_j$ significance(indicium$_j$), where summation is over indicia present in V
- Significance of an indicium is a function *f* of the importance of underlying *cphrs*
  - *f* must have monotonicity and concentration
  - $f(C) = \sum_{c \in C}$ exp(importance(c))
- Importance(c) = probability that c is hyperlinked to the corresponding Wikipedia article [M+07, M+09]

Outline

- Motivation
- Approach
- Experiments
- Concluding Remarks

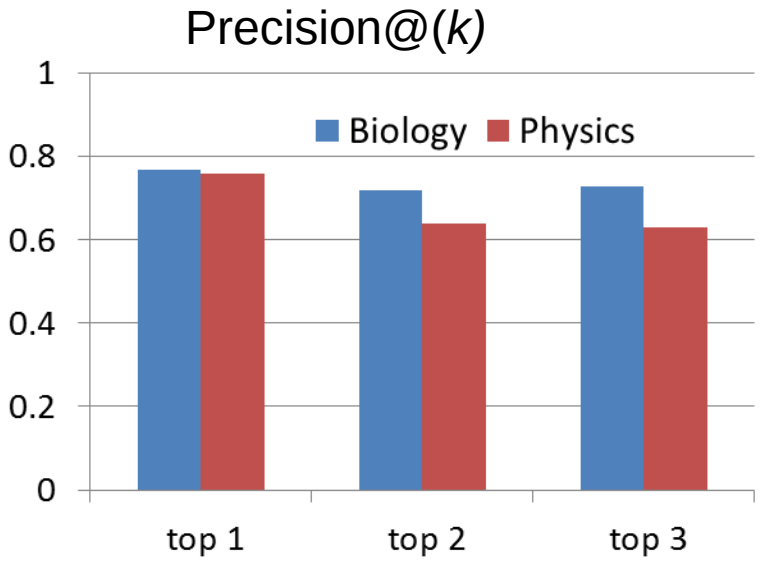
Data Sets

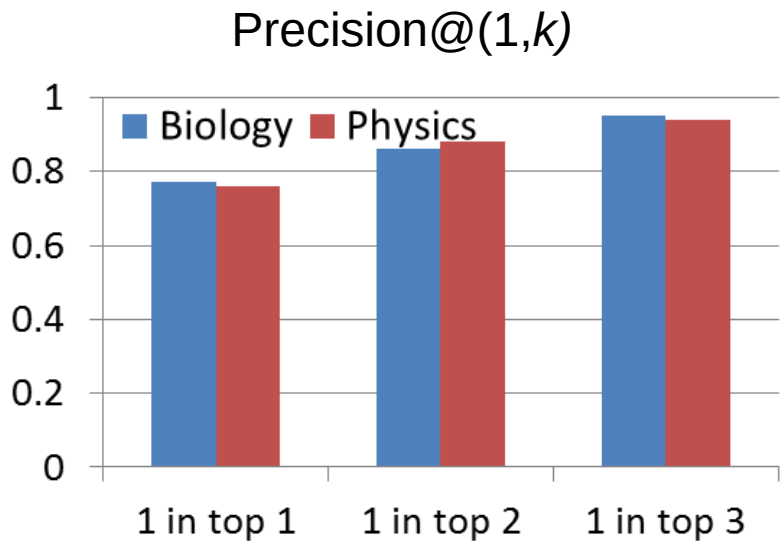
- Text corpus:
  - Books published by ck12.org
  - Books published by NCERT, India
- Video corpus:
  - 50,000 education-related, short videos obtained from focused crawl seeded with educational videos from a few reputed sites
  - Limited to videos that had accompanying user-uploaded transcript of the video content

Precision Study - Setup

- HIT: Read a section, watch a video, judge if the video is relevant (Default: Not Relevant)
- Each HIT judged by seven judges (Judge required to spend minimum 30 mins on a HIT)
- Precision@*k*: Fraction of videos in top *k* positions judged to be relevant
- Precision@(*i,k*): Whether at least *i* of the videos in top *k* positions were found relevant

Precision Study - Results



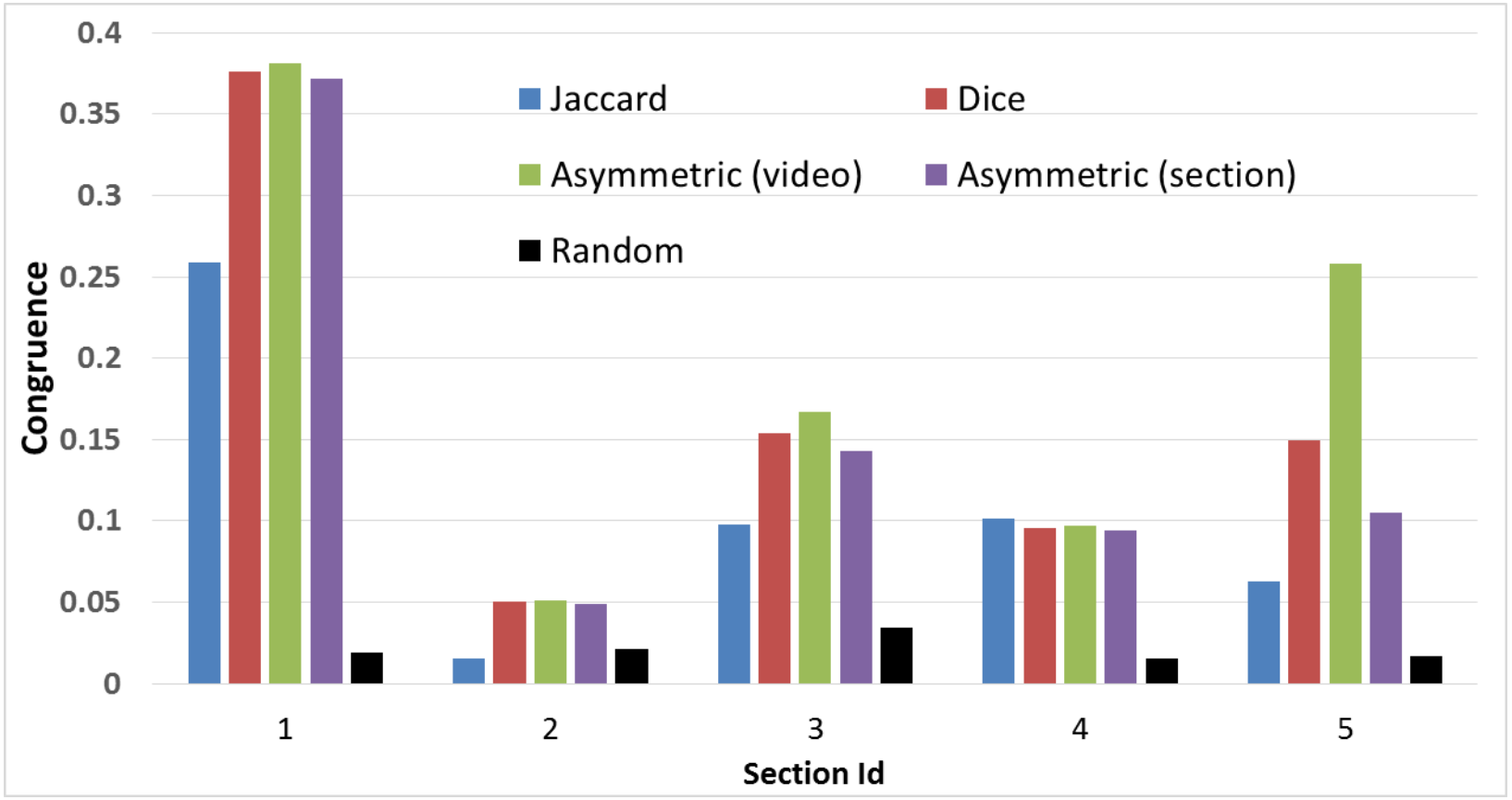


Congruence Study - Setup

- SectionHIT (videoHit): read section, provide 5 phrases that best describe the section (video)
- Each {section, video} pair given to 5 judges (Minimum 30 mins on a HIT)
- Convert phrases into unigrams (remove stop words)
- Performance metric: symmetric as well as asymmetric similarity measures wrt section and video (Jaccard, Dice)

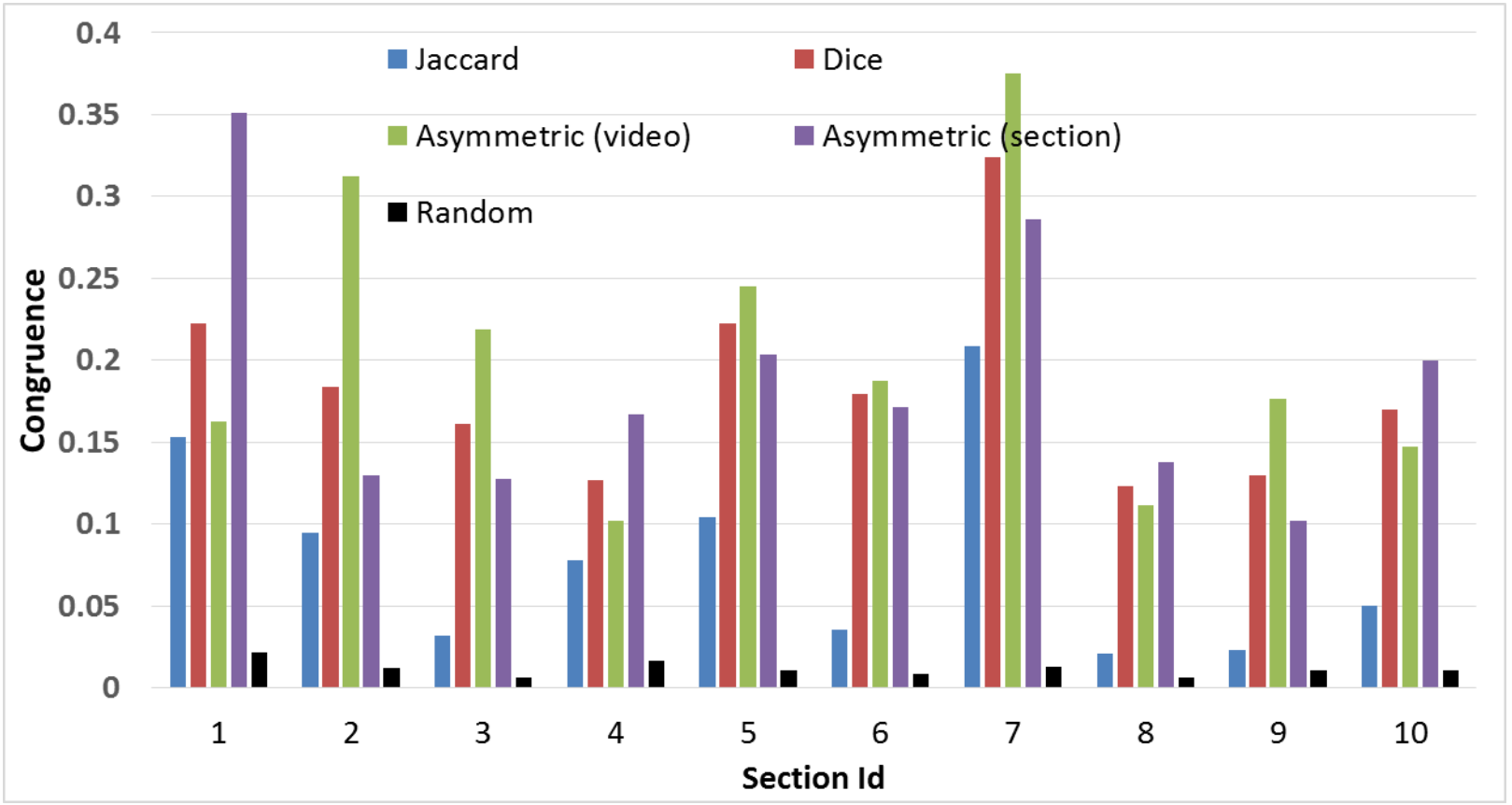
Congruence Study - Results

Physics Book



Congruence Study - Results

Biology Book



Illustrative Results

Magnetism 2

Khan Academy · 4,204 videos

Subscribe 1,781,710

239,035 views

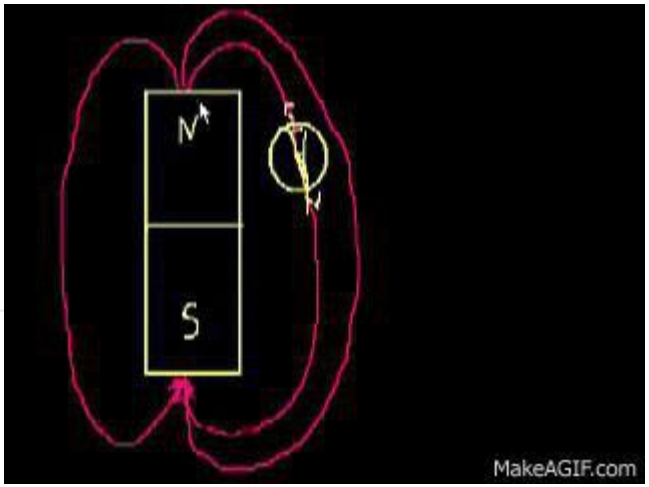
419 3

Like About Share Add to

Uploaded on Jul 30, 2008

More free lessons at: http://www.khanacademy.org/video?v=Nn...

Magnetic fields and their effects on moving electrical charges



<charged particle, dipole, field line>,
< charged particle, field line, magnetism, monopole >,…

"Can you feel a solar wind?" (Ask an Astronomer)

NASA Spitzer · 93 videos

Subscribe 6,350

2,823 views

12 1

Like About Share Add to

Uploaded on Jul 18, 2008

Dr. Robert Hurt explains what a solar wind is, and how it affects us here on Earth.

<solar wind, earth, magnetic field>,
<solar wind, earth, poles>,…

Microsoft Research

See Tech Report for examples from the Biology book

Outline

- Motivation
- Approach
- Experiments
- Concluding Remarks

Technical Report

- Inadequacy of alternate approaches to represent focus
  - TF/IDF
  - LDA
- Properties of scoring function
- More experimental results

Lessons Learnt

- Even widely used books might not follow principles of a well-written book
- An author's sectioning and paragraphing decisions might be arbitrary
- Concept structure in the transcripts can be important
- Evaluation methodology and performance metrics are not easy to come by
- Gap between theory and practice

References

[AGK+10] Rakesh Agrawal, Sreenivas Gollapudi, Krishnaram Kenthapadi, Nitish Srivastava, Raja Velu. "Enriching Textbooks Through Data Mining". DEV 2010.

[AGK+11a] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Identifying Enrichment Candidates in Textbooks". WWW 2011.

[AGK+11b] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Enriching Textbooks With Images". CIKM 2011.

[ACG+12] Rakesh Agrawal, Sunandan Chakraborty, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Empowering Authors to Diagnose Comprehension Burden in Textbooks". KDD 2012.

[AGK+13] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. "Studying from Electronic Textbooks". CIKM 2013.

[AJK14] Rakesh Agrawal, M. Hanif Jhaveri, and Krishnaram Kenthapadi. "Evaluating Educational Interventions at Scale". LAS 2014.

[ACG+14] Rakesh Agrawal, Maria Christoforaki, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi, Adith Swaminathan. "Augmenting Textbooks with Videos". ICFCA 2014.

[AGT14] Rakesh Agrawal, Behzad Golshan, Evimaria Terzi. "Grouping Students in Educational Settings". KDD 2014.