שיטות אוטומטיות להערכת איכות התוכן בוויקיפדיה

אתי יערי

אוניברסיטת בר אילן

yaariet@mail.biu.ac.il

הרצאה במסגרת יום העיון

"לא רק תחת הפנס: דרכים חדשניות לחיפוש ואחזור מידע איכותי "

כנס Info 2008, 1.4.08

"הרשת אף גרועה יותר מספריה מושחתת.
זאת, מכיוון שאלפי פיסות מידע לא מאורגנות נוספות מידי יום ביומו על ידי מספר עצום של תמהונים, מלומדים ואנשים, אשר יש בידם זמן, המשגרים את מסריהם הבלתי מסוננים אל המרחב הוירטואלי"

(Gorman, 1995, p. 34)

סייהולק (Ciolek):

WWW → MMM

MultiMedia Mediocrity

רשימות בדיקה (Checklists)

סמכות, דיוק, אובייקטיביות, עדכניות, כיסוי

- מיהו הגורם העומד מאחורי מקור המידע ומה המוניטין שלו ?
  האם זהו גוף ממשלתי/חינוכי או גוף מסחרי ?
- האם מצוינים מראי מקום לאימות המידע העובדתי ?
- האם מטרת המסמך מוצגת באופן ברור ?
- האם מצוין מתי המסמך נוצר ומתי עודכן לאחרונה ?
- האם נראה כי הנושא זכה לכיסוי באופן משביע רצון ?

- האם ישנו סרגל התמצאות לניווט פנימי באתר ?
- האם האתר דורש תוסף מיוחד לשם צפייה במידע?
  אם כן, האם מצורף קישור לדרכים שבהם ניתן להשיגו ?

הצלבה בין מקורות מידע שונים

- אישוש עצמי (Self-Sustaining)
- חוסר מוטיבציה לביצוע בדיקות ממושכות

- שיטת הערכה שונה שתתאים לתוכן המצוי בוויקיפדיה
- שיטה שהמשתמשים יאמצו

שיפוט אנושי: הערכת מומחים/מדרוג קוראים

- כמות חומר עצומה
- תדירות עדכון גבוהה

שיפוט אוטומטי: הערכה ממוכנת

הערכת טיבו של התוכן באופן אוטומטי,

על ידי ניתוח כמותי של היבטים שונים של המידע.

מטרת ההרצאה

להציג מספר דרכים שהוצעו בספרות ליישום הערכה ממוכנת של התוכן בויקיפדיה.

ניתוח קישורים פנימיים (1)

Deborah McGuinness ועמיתיה

(Knowledge Systems Laboratory, Stanford University)

$$\frac{\text{מספר הפעמים שהמילה מופיעה כקישור}}{\text{סך כל ההופעות של המילה במאגר}}$$

ככל שהתוצאה יותר גבוהה כך יותר ניתן לסמוך על הערך

0 מצביע על כך שלא ניתן לקבוע את מידת האמון

ניתוח קישורים פנימיים (2)

המילה "סיאטל" מופיעה 3855 פעמים
מתוכם 1408 פעמים כקישור ולכן:

1408/3855=0.36

בעיה:

מילים, שהסיכוי שלהם להופיע כקישור נמוך

אנגלית (0.003)
אהבה (0.004)
בירה (0.05)

MIT (0.52)
חוק גאוס (0.47)

ניתוח היסטורית שינויים (1)

* הוספה, מחיקה ושינוי הטקסט מצביעים על אמון

מידת האמון של גרסה מסוימת תלויה ב:

א. מידת האמון של הגרסה הקודמת

ב. מיהו התורם

ג. כמות הטקסט שמעורבת בשינוי

מידת האמון של ערך חדש נקבעת לפי המחבר

מפעיל מערכת / רשום / אנונימי / חסום

Trust Coloring

Luca de Alfaro ועמיתיו

University of California Santa Cruz (UCSC)

http://wiki-trust.cse.ucsc.edu/index.php/Main_Page

- טקסט שמופיע על גבי רקע לבן הוא טקסט, שניתן לבטוח בו באופן מלא.
- ככל שהרקע יותר כתום כך פחות ניתן לבטוח בו.

איך עובדת השיטה ? (1)

1. חישוב המוניטין של כל תורם.

- כאשר תורם מוסיף תוכן והתוכן שורד, המוניטין שלו עולה.
- כאשר תורם מוסיף תוכן, והתוכן נמחק/משוחזר (במהירות), המוניטין שלו יורד.

איך עובדת השיטה ? (2)

2. חישוב ערך ביטחון עבור כל מילה.

ערך הביטחון של מילה תלוי במוניטין של התורם שהוסיף אותה.

אם מילה שורדת עריכה היא זוכה למוניטין (גבוה יותר אם המילה קרובה לטקסט ששונה).

James G. Birney

James G. Birney

James Gillespie Birney (February 4, 1792–November 25, 1857) was an American presidential candidate for the Liberty Party in the 1840 and 1844 elections. He received 7,069 votes in the 1840 election and 62,273 votes in 1844.

James G. Birney was born in Danville, Kentucky. After studying at Transylvania College and Princeton, where he graduated in 1810, he studied law under Alexander J. Dallas in Philadelphia. He then began practice in Danville in 1814, and was elected to the State Legislature two years later. In 1818, Birney moved to the vicinity of Huntsville, Alabama. He had long opposed slavery, and had debated against it at Princeton, but was content with a gradual approach. While living in Alabama, he acted as agent for The National Colonization Society of America in 1832–33, which sought to send freed slaves to Liberia. In 1833, Birney returned to Kentucky, where he freed his own slaves. In 1839, he inherited 21 slaves from his father, all of whom he freed.

Birney by now had resolved that slavery should be brought to an immediate end. He organized the Kentucky Antislavery Society in 1835. Unable to find a publisher for an antislavery paper at Danville, he moved to Cincinnati, Ohio where he published the first issue of The Philanthropist on January 1, 1836. Hostile mobs destroyed his press several times over the next few years and Birney was himself repeatedly threatened.

Birney opposed all violence and supported the Constitution. He was elected secretary of the American Anti-Slavery Society in 1837. He gave many speeches before large assemblages of people, and became widely known as the leader of the Abolitionists who opposed violent or revolutionary measures. In 1845, he was disabled by a fall from his horse and spent the last twelve years of his life as an invalid.

His sons, William Birney (1819–1907) and David B. Birney (1825–64), both served as generals in the Union Army during the Civil War. His oldest son, James Birney, served as lieutenant-governor of the state of Michigan in 1860.

See also

- American Colonization Society
- Ohio History Central

sv:James G. Birney

Categories: United States presidential candidates | American abolitionists | 1792 births | 1857 deaths | Bleeding Kansas | People from Kentucky | People from Cincinnati

GNU FDL FREE DOC LICENSE

This page was last modified 17:25, 29 January 2007. This page has been accessed 3 times. Content is available under GNU Free Documentation License 1.2. Privacy policy About The UCSC Wikipedia Trust Project Disclaimers

Powered By MediaWiki

השוואת ארבעה מודלים

(Hu, M., Lim, E., Sun, A., Lauw, H. W., & Vuong, B.)

1. **מודל בסיסי –**
   ערכים בעלי איכות גבוהה הם אלו שנכתבו על ידי מחברים בעלי סמכות גבוהה (Authority); ומחברים בעלי סמכות גבוהה הם אלו שכותבים ערכים בעלי איכות גבוהה

2. **מודל ביקורת עמיתים –**
   טקסט ששרד לאחר עריכה שביצע מחבר בעל סמכות גבוהה הוא בעל איכות גבוהה

3. **מודל הסתברות הביקורת –**
   ניתן להחיל את עקרונות מודל ביקורת עמיתים רק על מילים הסמוכות לטקסט שעבר עריכה

4. **מודל נאיבי –**
   אורך ערך מרמז על איכות. יתר ארוך יותר איכותי

0011

מידת הישרדותו של טקסט

טקסט ששורד = טקסט שניתן לסמוך עליו

בדיקה על הערכים שנבחנו במחקר של Nature:

20% מן הטעויות ניתן לייחס לטקסט שנכתב בעריכה ראשונה ושרד את כל העריכות

Luyt, Aaron, Thian & Hong (JASIST, January 2008)

↓

First-mover Advantage

תוכן, אשר נוסף בעריכות הראשונות נוטה לשרוד לאורך זמן בלא קשר לאיכותו ולנכונותו.

הערכה ממוכנת על בסיס תובנות משתמשים

לאילו נימוקים ניתן להצמיד ערך מספרי ללא התערבות אנושית ?

א. נימוק שאינו בר מדידה אוטומטית (מדידה סובייקטיבית)

ב. נימוק בר מדידה אוטומטית (מדידה אובייקטיבית)

מדידה ≠ משמעות

- נוכחות של מאפיין
- כמות ממאפיין

"מדדים פשוטים"

- כמות עריכות ביחס לגיל הערך
- מידת גיוון (מספר העורכים הייחודיים)
- אורך ערך (בשילוב עם כמות העריכות)
- האם קיימים בערך קישורים חיצוניים
- פעילות בדף השיחה (אורך, מספר משתתפים)

תודה על ההקשבה !

שאלות ?

0011