****

**מחלקת הנדסת תוכנה**

שם הפרויקט: סוכן רשת אוטומטי

Project Name: Automatic Web Agent

דו"ח ביניים 1

|  |  |
| --- | --- |
| שם הסטודנט: | קובי הרשקוביץ |
| מספר תעודת זהות: | 025632654 |
| שם המנחה: | אלעד הוגן |
| תאריך ההגשה: | 8/9/09 |

**תוכן ענינים**

[[1 תקציר 4](#_Toc229036130)](#_Toc240163012)

[[2 סקירה ספרותית 4](#_Toc229036130)](#_Toc240163013)

[[3 מטרות העבודה 6](#_Toc229036130)](#_Toc240163014)

[[4 סקירת מצב קיים 6](#_Toc229036130)](#_Toc240163015)

[[5 ניתוח חלופות מערכתי 7](#_Toc229036130)](#_Toc240163016)

[[6 ניתוח חלופות טכנולוגיות 8](#_Toc229036130)](#_Toc240163017)

[[7 הפתרון שנבחר 8](#_Toc229036130)](#_Toc240163018)

[[8 תכנית בדיקות 9](#_Toc229036130)](#_Toc240163019)

[[9 תוכנית עבודה של הפרויקט 10](#_Toc229036130)](#_Toc240163020)

[[10 ניהול סיכונים 11](#_Toc229036130)](#_Toc240163021)

[[11 רשימת מקורות: 11](#_Toc229036130)](#_Toc240163022)

[[נספח ב1 - (Software Requirements Document (SRD 13](#_Toc229036130)](#_Toc240163023)

[[נספח ב2 - (Software Test Plan (STP 25](#_Toc229036130)](#_Toc240163024)

[[נספח ב3 - (Software Project Management Plan (SPMP 27](#_Toc229036130)](#_Toc240163025)

[[נספח 1 – הצעת הפרויקט 28](#_Toc229036130)](#_Toc240163026)

# תקציר

כיום השימוש באינטרנט נעשה רווח כמעט בכל בית ובאמצעותו ניתן לבצע משימות רבות ומגוונות החל מקבלת מידע בידורי וכלה בפעולות בנקאיות. במרבית מן הפעולות נדרש המשתמש להזדהות, לבצע מעקב אחר נתונים באתר, ולבצע פעולות על סמך ידע שלמד מאתרים אחרים. פעולות אלו במרבית הזמן זהות/דומות בין האתרים וכמו כן ביצוען עלול לדרוש תשומת לב וזמן רב מן המשתמש.

הפרוייקט המוצע הינו סוכן רשת אוטומטי הבא להחליף מטלות שגרתיות וגוזלות זמן של משתמש דפדפן ממוצע/מתקדם. באמצעות המערכת המוצעת יוכל המשתמש לתכנן תסריט של פעולות מסויימות באתרים מסויימים בהתאם למידע הנדגם מאתרים אחרים ולקבוע את זמן ביצוע התסריט.

לדוגמא ניתן לקבוע תסריט שבו נדגם כל 4 שעות אתר סקי שהמשתמש מחבב במיוחד (לדוגמא www.champery.com) , התסריט מכוון לזהות את השדה של גובה השלד באתר וברגע שגובה השלג מגיע ל-110cm התסריט שולף מאתר ה-"דקה ה-90" את מחירי חבילות הנופש לאתר ושולח דוח מפורט עם הנתונים באי-מייל למשתמש.

דוגמא נוספת יכולה להיות תסריט שיאריך את זמן השאלת ספר באתר הספריה של מכללת אפקה ,שאינו מצוייד כיום בסוכן אוטומטי להארכת השאלות, פעם בשלושה ימים.

יחודיות הפרוייקט היא שבניגוד לסוכנים אוטומטיים הנמצאים באתר מסויים ונשלטים מתוך האתר הפרוייקט הנ"ל מציע סוכן שיתוכנת בצד הלקוח ולא בצד השרת, דבר אשר יאפשר לו נוחויות ושליטה מירבית להשגת מטרותיו.

לחילופין במישור העסקי, חברה לממכר מוצרים באיטרנט יכולה לבצע מעקב אחר מוצרים מקבילים באתרי מתחרים ולעדכן את המחירים באתר שלה ע"פ התוצאות הנדגמות.

# סקירה ספרותית

1. סוכני רשת הם מערכות תוכנה מורכבות אשר פועלות ברשת האינטרנט העולמית (www), ברשת הארגונית או ברשתות מקומיות. סוכני רשת מתוכננים לבצע מגוון משימות החל מאגירה וניתוב של מידע וכלה בחיפוש קיטלוג וסינון מידע.

הספרות מתארת מסגרת לתכנון וניתוח של סוכני רשת ומערכות סוכנים בהתבסס על מודלים מתמטיים של הסביבה בה הם פועלים. יש 3 אבני בניין מרכזיות לבניית סוכני רשת:

* מודלים סטוקאסטיים של גרף הרשת אשר משקללים את הפילוג והקישוריות של דפי רשת ונותנים קוים מנחים מרכזיים בתכנון סוכנים, ע"י הפקת תכונות ייחודיות של הסביבה בה הסוכנים פועלים.

הרשת יכולה להיראות כגרף מכוון אשר הצמתים בו הם דפי HTML סטטיים והקשתות בו הם קישורים מאתר אחד למשנהו. מאמר [ ‎18] מציע ומנתח מודלי גרפיים אקראיים שנוצרו ע"י תצפיות נסיוניות ברשת. המודלים שונים מהמודלים המקובלים בכך ש קשתות שנבחרו בצורה בלתי תלויה לא משתקפות בסטטיסטיקה שנצפתה ברשת, כך שקשתות במודלים אלו הינם תלויות סטטיסטית ביניהן. כמו כן נוספים צמתים לגרף כתלות בזמן. דבר זה משקף את העובדה שהרשת משתנה עם הזמן.

* שיטות למידה סטטיסטית המאפשרות לסוכנים ללמוד על סביבתם ע"י הפקת מודלים סטוקאסטיים של דף רשת ומבנה קישורים מקומיים. מאמר [ ] מתאר את הרשת כעולם שבו המוני קישורים מחברים בין מסמכי ודפי אתרים כך שנוצר גרף בעל סיבוכיות גבוהה וללא תקדים וזהו גרף הרשת. מאמר זה מציג גישה חדשה ללימוד מודלים הסתברותיים שך הרשת אשר יכולים לשמש כחלק בתהליך חיזוי קישוריות ומידע של מסמכי רשת. השיטה המתוארת במאמר זה היא שיטה הסתברותית לצמצום מימדים (probabilistic dimension reduction) אשר גוזרת ומאחדת את שיטת הניתוח הסמנטי המאוחר (Latent Semantic Analysis) ואת אלגוריתם קליינברג (Hubs-and-Authorities) בסביבה סטטיסטית.

דבר זה אמור להיות הצעד הראשון לפיתוח תשתית סטטיסטית של טכנולוגיות מידע ברשת. למרות שמאמר זה לא מתמקד ביישום מסויים, מגוון אלגוריתמים הפועלים בסביבת רשת יכולים לנצל את היתרונות בטכניקות המוצגות למטרות שונות כגון מנועי חיפוש, "זוחלי רשת" (Web crawlers) ומערכות סוכני מידע.

* אלגוריתמים לתכנון אוטונומי וקבלת החלטות בסביבת רשת לאפשר לסוכנים להשיג מטרות ולהסתגל לסביבה משתנה [16]

ניקח לדוגמא בעיות קלאסיות בתחום בסיסי נתונים/איחזור מידע כגון חיפוש בטקסט, כריית מידע וסיווג. הקלט לבעיות מסוג זה הוא בדרך כלל אוסף של נתונים/מסמכים. הרשת עם המבנה הנוסף שלה כגרף, מאפשרת את השיפור של טכניקות קיימות עם טכניקות מעולם הגרפים. במאמר הודגם שיפור זה באמצעות פתרונות מבוססי גרפים לבעיות הבאות: חיפש נושא, מיספור נושא, סיווג וזחילה (crawling).

1. ככל שאפליקציות רשת נעשו מסובכות יותר, בדיקתן נעשתה יותר חשובה. יש מספר רב של טכניקות בדיקה הנגישות למתכנת, לדוגמא במהדורת אפריל 2005 של *MSDN*®*Magazine* יש תאור של מערכת בדיקה מבוססת Jscript שיכולה לבדוק אפליקציית רשת באמצעות ממשק המשתמש שלה על-ידי שימוש במודל האובייקטים של IE (DOM) טכניקה זו יעילה אך בעךת מספר מגבלות. במאמר זה מתוארת טכניקה בה משתמשים ב-DOTNET ליצירת בדיקת אפליקציית רשת כאשר הספריות המשמשות ברמה הנמוכה הן fmshtml.dll ו- shdocvw.dll שתפקידן לגשת ולשנות את ה-DOM.[20]

# מטרות העבודה

**מטרות:**

רבות מהפעולות הפיננסיות, הלימודיות והבידוריות נעשות באמצעות דפדפן ברשת האינטרנט:

* רישום לאתרים
* קניה/מכירה של מניות
* הזמנת כרטיסים
* הגשת מטלות
* בחירת קורסים

רבות ממטלות אלו הן מחזוריות, שגרתיות ועלולות להתיש את המשתמש ו/או לגרום לו לבצע שגיאות באופן הכנסת הנתונים.

מטרת הפרוייקט הנגזרת מבעיה זו היא בניית אפליקציה אשר תבצע מטלות שגרתיות וצפויות מראש בצורה אוטומטית במקום משתמש דפדפן אנושי

**יעדים:**

ניתן לפרוט מטרת על זו ל-3 יעדים:

* ניטור אתר בצורה מחזורית
* ביצוע פעולות אוטומטיות בדפדפן
* התראה למשתמש ע"פ בקשתו על אירוע באתר מסויים.

**מדדים:**

המדדים ע"פ תימדד הצלחת הפרוייקט הם:

* אחוז האתרים הנתמכים מתוך עשרת האתרים הראשונים בחיפוש בגוגל יעמוד על 80% בקטגוריות פיננסים, בידור, ספורט (דוגמא לחיפוש יכולה להיות המילה 'פיננסי')
* נוהל כניסה לאתר הספרייה במכללה והארכת השאלה אם קיימים ספרים מושאלים יסתיים בהצלחה.
* נוהל דגימת אתר לשינויים בתדירות של 6 פעמים בדקה יסתיים בהצלחה (ידווח בצורה נכונה על שינוי).
* משתמש חדש יוכל להגדיר תסריט אוטומטי של דגימת 2 אתרים ואינטראקציה ביניהם בתוך כשעה.

# סקירת מצב קיים

כיום כאשר משתמש רוצה לבצע מעקב/שינוי נתונים באתר מסויים פרושות לפניו מספר אפשרויות:

תמיכה של האתר עצמו בסוכן רשת אשר מיידע את המשתמש לגבי נתון ייחודי לאתר אליו הוא שייך. לדוגמא האתר ebay מציע סוכן שיציע במכרז מחיר עבור פריט מסויים עד תקרה שיקבע המשתמש.

מעקב מחזורי אישי של המשתמש אחר שינויים באתר.

שימוש בתוכנות אוטומציה יקרות כדוגמת QFT או RFT.

QTP(Quick Test Pro ) היא חבילת תוכנה מבית מרקורי שבה ניתן לבצע אוטומציה לאפליקציות חלונאיות ורשתיות כגון אוטומציה לכתיבת דואר אלקטרוני ושליחתו או חקירת דף רשת והכנסת נתונים אליו.

RFT (Rational Functional Tester) היא גם חבילת תוכנה לבדיקות אוטומטיות מבית IBM המאפשרת גם לבצע בדיקות על אלמנטים מסוג Flex וטרמינלים כגון VT102.

מחירם של כל אחת מחיבולות אלו מגיע לכמה אלפי דולרים עבור רשיון למשתמש אחד.

# ניתוח חלופות מערכתי

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| חלופה | קריטריון להשוואה | יתרון | חסרון | משקל (חשיבות החלופה יחסית לכלל הפרוייקט) |
| 1. הקלטת אלמנטים עבור בסיס הנתונים באמצעות WebControl | ייצור בסיס הנתונים עבור האלמנטים של האתרים . | קלות בבחירת האלמנטים (WYSIWYG) | סיבוכיות בתכנות ממשק המשתמש | 30% |
| 1. הכנסת אלמנטים לבסיס הנתונים באמצעות טופס מילוי | קלות בתכנות צורת קלט זו | יש צורך במומחיות מסויימת של המשתמש וביכולתו לנתח את הדף.(לדוגמא באמצעות view source) |
| 1. כתיבת תסריט באמצעות שפת תסריטים ומפרש | תסריט פעולה | גמישות בפעולות האפשריות לביצוע על אתרים | על המשתמש ללמוד שפת תסריט, כמו כן נוצרת סיבוכיות של יצירת מפרש (תכנות מורכב) | 50% |
| 1. ייצור התסריט באמצעות טופס מובנה ובקרים | קלות בייצור התסריט מבחינת המשתמש | מגוון פעולות מוגבל יחסית לחלופת התסריט החופשי. |
| 1. עבודה מול דפדפן מחוץ לאפליקציה | ממשק משתמש | זהו עולם מוכר למשתמש | קושי בהתממשקות לדפדפן חיצוני | 20% |
| 1. עבודה מול דפדפן המשובץ בתוך האפליקציה |  |  |

טבלה 1

**סיכום**: בכדי לעמוד בלוח הזמנים תוך כדי שמירה על איכות נבחרו הרכיבים הבאים מבין החלופות – 1,4,6.

חלופה 1 נבחרה מכיוון שכבר ערכתי ניסוי לבחירה אינטראקטיבית של אלמנט מדף והתרשמתי שיהיה יחסית פשוט למימוש בצורה זו.

חלופה 4 נבחרה מכיוון שהסיבוכיות ביצירת שפת תסריט ומפרש לה מהווה סיכון מבחינת הצלחת המשימה ועמידה בזמנים.

חלופה 6 נבחרה משום שבדפדפן מובנה יש שליטה רבה יותר כמו כן כך המשתמש יידע בוודאות מהם מרכיבי האפליקציה.

# ניתוח חלופות טכנולוגיות

כדי לממש את המערכת יש צורך במספר טכנולוגיות שונות:

* שפת התכנות בה ייכתב הקוד – הן מבחינת יעילות השפה הן מבחינת קלות הכתיבה בה והן מבחינת עקומת הלימוד.
* סביבת הפיתוח
* כיצד ובאיזה מבנה יישמרו התסריטים/אלמנטי האתר – סיבוכיות המימוש, יעילות הגישה לנתונים.
* ספריית הגישה למודל האובייקטים בדף (DOM Library)

בטבלה הבאה מרוכזים החלופות השונות , הציון הניתן לכל חלופה הוא מ-1 עד 10 כאשר 10 הוא הטוב ביותר.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| נושא לחלופה | חלופה 1 | | | | חלופה 2 | | | | חלופה 3 | | | |
| שם | יתרונות | חסרונות | ציון | שם | יתרונות | חסרונות | ציון | שם | יתרונות | חסרונות | ציון |
| שפת תכנות  משקל:25% | CPP | יעילה, רוב ספריות API כתובות בה | מסובכת יחסית | 7 | JAVA | פורטבילית, קל לכתוב בה | ספריה גרפית מגושמת, פחות יעילה | 6 | C# | המוכרת לי ביותר, כתיבה פשוטה, ספריות גרפיקה ותקשורת עשירות | נתמכת בעיקר בסביבת windows | 9 |
| סביבת עבודה  משקל:15% | MS Visual Studio | תעשייתית, ממשק משתמש נוח, אמינה | עלות יקרה | 9 | Eclipse | חינם, קהילה רחבה של משתמשים | מותאמת בעיקר ל-JAVA | 8 | CodeBlocks | חינם | סביבה דלה במשתמשים מותאמת בעיקר ל-CPP | 5 |
| ספריית DOM Library  משקל:40% | Native (mshtml.dll shdocvw.dll) | שליטה מלאה בכל מה שניתן לעשות ב-DOM | כתובה ב-CPP שפה יותר מסובכת עבורי, ללא שליטה בדיאלוגים | 8 | Watin | מבוססת C# קלה לשליטה ובעלת תמיכה בדיאלוגים | לא תומכת בכל האפשריות של ה-DOM | 8 | Selenium | מבוססת JAVA פשוטה, תמיכה נרחבת | לא קל לשלב אותה באפליקציה חלונאית | 6 |
| מבנה נתונים משקל: 20% | MS SQL | תעשייתי, מתממשק ל-C# מיידית | מסובך | 8 | SQLITE | פשוט, קל משקל, מוכר, חינמי | יש תמיכה רק בחלק מ-SQL סטנדרטי | 9 | קבצי MS SQL | חינמי, תמיכה רחבה. | שמירה מקומית, הצפנה | 9 |

# הפתרון שנבחר

לאחר שיקולי כל הגורמים על פי שקלול הציונים והמשקלות בטבלה הנ"ל נעשתה הבחירה הבאה:

* שפת תכנות: C# - השפה המוכרת לי ביותר, יש בה שילוב של עוצמה בעיצוב מהיר של אפליקציות חלונאיות ביחד עם WebControl נוח לשליטה.
* סביבת עבודה: Visual Studio 2008 – סביבה זמינה במכללה.
* ספריית DOM Library: Watin – ספריית קוד פתוח המפשטת טיפול בדיאלוגים ואלמנטים ואשר תחסוך זמן תכנות רב , כך שישאר יותר זמן לעסוק בלוגיקת המערכת מאשר בפרטים טכניים.
* מבנה נתונים: בסיס הנתונים MS SQL מבוסס קובץ (ללא שרת) – בסיס נתונים אמין בעל הצפנה מובנית, תואם לסביבת הפיתוח, פשוט ובעיקר חינמי.

למערכת יהיו שלושה חלקים עיקריים ממשק הקלטת השדות הנדגמים, מודול ייצור התסריטים ומודול התזמון. דרך מימושם מתוארת בנספח ה-SRD (נספח ב1).

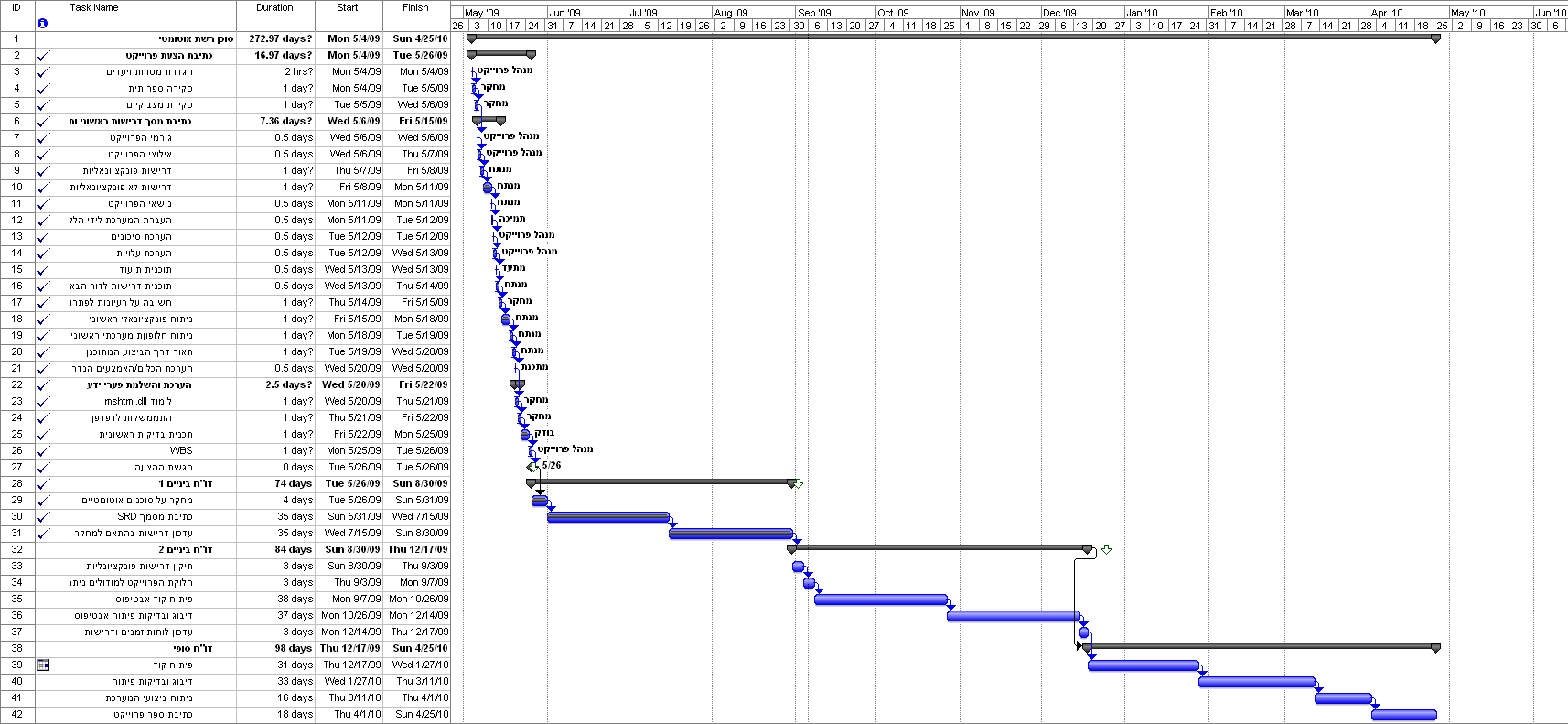
# תכנית בדיקות

סביבת הבדיקה הבסיסית תכיל 3 אתרי דמה בעלי מגוון פקדים, הבדיקות השונות יכללו נסיון להקליט פקדים ושדות מאתרים אלו, ליצור מספר תסריטים שיכללו את מגוון האפשרויות הניתנות לתסרוט, והפעלתם בתזמון מתוכנן מראש.

לתוכנית המפורטת ראה נספח ב2.

# תוכנית עבודה של הפרויקט

על פי התוכנית שהוגשה בהצעה (נספח 1)



# ניהול סיכונים

|  |  |  |  |
| --- | --- | --- | --- |
| סיכון | סיכוי | השפעה | דרך מניעה |
| חריגה מלוחות זמנים | 50% | אי הגעה לאבני דרך בפרוייקט | ניסיון להקדים ביצוע של משימות והקדשת זמן נוסף לפרוייקט |
| כשל חומרתי בסביבת הפיתוח/ריצה | 20% | עיכוב של עד כשבועיים | גיבוי כל איטרציה במערכת ניהול קוד |
| אי יכולת לזהות אלמנטים מיוחדים בדף | 50% | אי יכולת להשתמש באלמנטים אלו – ירידה בפונקציונאליות המובטחת | עדכון סעיף האילוצים בשלב מוקדם ככל האפשר |
| דפים דינמיים (Ajax, DHTML) | 80% | שיבוש יכולת זיהוי האלמנטים | התקנת מנגנוני המתנה לאלמנטים לא קיימים. |

# רשימת מקורות:

1. Jon Kleinberg. **Authoritative sources in a hyperlinked environment**. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997. [[pdf]](http://www.cs.cornell.edu/home/kleinber/auth.pdf)
2. Soumen Chakrabarti, Martin van den Berg, Byron Dom, **Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery**, WWW8 [[html]](http://www8.org/w8-papers/5a-search-query/crawling/index.html)
3. Steve Lawrence and Lee Giles, **Searching the World Wide Web,** Science 1998 [[pdf]](http://www.neci.nj.nec.com/~lawrence/papers/search-science98/search-science98.pdf)
4. Jeffrey Dean, Monika R. Henzinger, **Finding Related Pages in the World Wide Web**, WWW8 [[html]](http://www8.org/w8-papers/4a-search-mining/finding/finding.html)
5. Oren Zamir and Oren Etzioni, **Grouper: A Dynamic Clustering Interface to Web Search Results**, WWW8, [[html]](http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html)
6. Rajan M. Lukose and Bernardo A. Huberman, **Surfing as a Real Option** [[abstract]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/options.html)
7. Bernardo A. Huberman and Rajan M. Lukose, **Social Dilemmas and Internet Congestion** [[abstract]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/storms.htm)
8. Bernardo A. Huberman, Peter L.T. Pirolli, James E. Pitkow, and Rajan M. Lukose, **Strong Regularities in World Wide Web Surfing**, Nature ??? [[abstact]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/surfing.html)
9. William W. Cohen, Wei Fan, **Web-Collaborative Filtering: Recommending Music By Spidering the Web**, WWW9 [[html]](http://www9.org/w9cdrom/266/266.html)
10. Mike Perkowitz Oren Etzioni, **Towards Adaptive Web Sites: Conceptual Framework and Case Study**, WWW8 [[html]](http://www8.org/w8-papers/2b-customizing/towards/towards.html)
11. Marc Langheinrich, Atsuyoshi Nakamura, Naoki Abe, Tomonari Kamba, Yoshiyuki Koseki, **Unintrusive Customization Techniques for Web Advertising**, WWW8 [[html]](http://www8.org/w8-papers/2b-customizing/unintrusive/unintrusive.html)
12. Jim Hendler, **Is there an intelligent agent in your future?** Nature, Web matters, 11 March 1999. [[html]](http://helix.nature.com/webmatters/agents/agents.html)
13. Steve Kirsch, **The future of Internet search (keynote address)**, SIGIR 1999 [[pdf]](http://www.acm.org/pubs/articles/proceedings/ir/290941/p1-kirsch/p1-kirsch.pdf)
14. G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure, Proceedings of the 8th International Computing and Combinatorics Conference (COCOON), 2002.
15. [G. Pandurangan](http://www.cs.brown.edu/people/gopal) , [P. Raghavan](http://theory.stanford.edu/people/raghavan/index.html) , and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**Building Low-Diameter P2P Networks**](http://www.cs.brown.edu/research/webagent/focs-2001.pdf). Proceedings of the 42th IEEE Symp. on Foundations of Computer Science. 2001.
16. S.R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**The Web as a graph**](http://www.cs.brown.edu/research/webagent/pods-2000.pdf). Proceedings of the 19th ACM Symposium on Principles of Database Systems, pp 1-10, 2000.
17. Thomas Hofmann, [**Learning Probabilistic Models of the Web**](http://www.cs.brown.edu/research/webagent/Hofmann-SIGIR00.pdf), ACM SIGIR 2000
18. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**Stochastic models for the Web graph**](http://www.cs.brown.edu/research/webagent/focs-2000.pdf). Proceedings of the 41th IEEE Symp. on Foundations of Computer Science. 2000.
19. <http://en.wikipedia.org/wiki/Trident_%28layout_engine%29> , downloaded 1/9/2009
20. <http://msdn.microsoft.com/en-us/magazine/cc163723.aspx> , downloaded 1/9/2009

# נספח ב1 - (Software Requirements Document (SRD

**(מבוסס על IEEE Std 830-1998)**

**1 Introduction**

**1.1 Purpose.**

The purpose of this document is to describe the requirement needed for the Automatic Web Agent System system.

The document is divided to sections, where as each section will describe the requirements from a single module.

**1.2 Scope of the software**

Nowadays web use has become common to practically every household. The web enables to perform many diverse tasks, ranging from acquiring entertainment data (movies, songs, gossip etc…) to financial transactions.

Most of the tasks a user might perform, require the user to authenticate his identity, track data on the website and perform actions based on knowledge acquired on other websites. These actions are very similar in most of the tasks regardless of their nature, all of which require the attention of the user and might be extremely time consuming.

The Automatic Web Agent application project was devised to relieve the average/advanced user from these routine and time consuming tasks.

The Automatic Web Agent will enable the user to compose a script of actions to be taken on specific websites according to data sampled on other sites and conditions that the users will set, and determine the scheduling of the script execution.

For example, there can be a script that will sample the users favorite sky website every 4 hours and will monitor the snow depth of the sky track. Once a preset depth is detected the script will extract the prices of plane tickets to the sky track and email them to the user.

Another example can be a script that will extend the lending period of a book in the Afeka college library, which is not equipped today with a built-in agent, once every 3 days.

Alternatively, an example from the business sector could be a for a company which sells products via the web, a script can track the competitors web sites for price changes and update the prices of the products accordingly.

**1.3 Definitions, acronyms and abbreviations.**

|  |  |
| --- | --- |
| DOM | Document Object Model |
| WWW | World Wide Web |
| DB | Data Base |
| Site Object | a C# Object that represent a specific element in the DOM of the website |
| HTML |  |
| DHTML |  |
| URL |  |
| HTTP |  |
| SQL |  |
|  |  |

Table

**1.4 References**.

1. Jon Kleinberg. **Authoritative sources in a hyperlinked environment**. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997. [[pdf]](http://www.cs.cornell.edu/home/kleinber/auth.pdf)
2. Soumen Chakrabarti, Martin van den Berg, Byron Dom, **Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery**, WWW8 [[html]](http://www8.org/w8-papers/5a-search-query/crawling/index.html)
3. Steve Lawrence and Lee Giles, **Searching the World Wide Web,** Science 1998 [[pdf]](http://www.neci.nj.nec.com/~lawrence/papers/search-science98/search-science98.pdf)
4. Jeffrey Dean, Monika R. Henzinger, **Finding Related Pages in the World Wide Web**, WWW8 [[html]](http://www8.org/w8-papers/4a-search-mining/finding/finding.html)
5. Oren Zamir and Oren Etzioni, **Grouper: A Dynamic Clustering Interface to Web Search Results**, WWW8, [[html]](http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html)
6. Rajan M. Lukose and Bernardo A. Huberman, **Surfing as a Real Option** [[abstract]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/options.html)
7. Bernardo A. Huberman and Rajan M. Lukose, **Social Dilemmas and Internet Congestion** [[abstract]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/storms.htm)
8. Bernardo A. Huberman, Peter L.T. Pirolli, James E. Pitkow, and Rajan M. Lukose, **Strong Regularities in World Wide Web Surfing**, Nature ??? [[abstact]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/surfing.html)
9. William W. Cohen, Wei Fan, **Web-Collaborative Filtering: Recommending Music By Spidering the Web**, WWW9 [[html]](http://www9.org/w9cdrom/266/266.html)
10. Mike Perkowitz Oren Etzioni, **Towards Adaptive Web Sites: Conceptual Framework and Case Study**, WWW8 [[html]](http://www8.org/w8-papers/2b-customizing/towards/towards.html)
11. Marc Langheinrich, Atsuyoshi Nakamura, Naoki Abe, Tomonari Kamba, Yoshiyuki Koseki, **Unintrusive Customization Techniques for Web Advertising**, WWW8 [[html]](http://www8.org/w8-papers/2b-customizing/unintrusive/unintrusive.html)
12. Jim Hendler, **Is there an intelligent agent in your future?** Nature, Web matters, 11 March 1999. [[html]](http://helix.nature.com/webmatters/agents/agents.html)
13. Steve Kirsch, **The future of Internet search (keynote address)**, SIGIR 1999 [[pdf]](http://www.acm.org/pubs/articles/proceedings/ir/290941/p1-kirsch/p1-kirsch.pdf)
14. G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure, Proceedings of the 8th International Computing and Combinatorics Conference (COCOON), 2002.
15. [G. Pandurangan](http://www.cs.brown.edu/people/gopal) , [P. Raghavan](http://theory.stanford.edu/people/raghavan/index.html) , and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**Building Low-Diameter P2P Networks**](http://www.cs.brown.edu/research/webagent/focs-2001.pdf). Proceedings of the 42th IEEE Symp. on Foundations of Computer Science. 2001.
16. S.R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**The Web as a graph**](http://www.cs.brown.edu/research/webagent/pods-2000.pdf). Proceedings of the 19th ACM Symposium on Principles of Database Systems, pp 1-10, 2000.
17. Thomas Hofmann, [**Learning Probabilistic Models of the Web**](http://www.cs.brown.edu/research/webagent/Hofmann-SIGIR00.pdf), ACM SIGIR 2000
18. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**Stochastic models for the Web graph**](http://www.cs.brown.edu/research/webagent/focs-2000.pdf). Proceedings of the 41th IEEE Symp. on Foundations of Computer Science. 2000.
19. <http://en.wikipedia.org/wiki/Trident_%28layout_engine%29> , downloaded 1/9/2009
20. <http://msdn.microsoft.com/en-us/magazine/cc163723.aspx> , downloaded 1/9/2009

**1.5 Overview of the document.**

This document is mainly aimed for the developers and technical staff involved in the project, and will provide a basic understanding of the requirements to be met on project completion. The main points of interest for a developer would be:

1. The functional requirement listing, in section 3, interface requirements and the security and performance requirements near the end of that section.

2. The model description in section 2.7.

For the IT personnel, the primary points of interest are:

1. Environment considerations – section 2.4

2. Resource requirements – sections 3.4, and

3. Security requirements – section 3.6

**2 General Descriptions**

**2.1 Relation to current projects.**

This project is somewhat dependent on the open source Web Test API called WatiN [[1]](#footnote-2) the current version used (WatiN-2.0.10.928-net-2.0) should be enough for this project.

**2.2 Relation to predecessor and successor projects.**

NA – this is a new project.

**2.3 Function and purpose**

|  |  |
| --- | --- |
| Function | purpose |
| DOM element recording | allow a method for the user to populate the DB with mappings between DOM elements and Site Objects |
| Site Objects management | allow user to change recognition parameters of a DOM element contained in a Site Object |
| Script composition and validation | Allow the user to create the automation script and validate it. |
| Script management | Allow the user to rename, save or delete scripts |
| Script scheduling | Allow the user to schedule the execution time and frequency of the script. |
| Script Execution | Actual execution of the actions in the scripts which include monitoring and manipulating websites and notifications to the user. |

Table

**2.4 Environmental considerations.**

The product will be used on the customer PC as a managed windows application.

The target audience of this product is intermediate to advanced browser user with beginner knowledge in scripting concepts.

The product is designed to run on:

* X86 architecture
* 1GB RAM / HD space TBD
* windows XP OS.

**2.5 Relation to other systems.**

Any website is hosted on a hosting server (HTTP server)

HTTP Server

Auto Web Agent application

WWW

Context diagram 1

When a browser navigates to a website it actually sends a request for a page from the HTTP server (via the URL entry in the browser which translates to a HTTP request). The server then sends the html document to the browser which constructs from it a DOM tree and then renders it to the screen.

In this project the browser is embedded in the Auto Web Agent application.

**2.6 General constraints.**

* Only pure DHTML web sites will be supported (i.e. no flash/flex/SilverLight/activex support).
* The project provided is intended as a proof-of-concept only
* The product will be graphically designed to some extent but the development effort will be put into functionality and features as a priority.

**2.7 Model description.**

The following diagrams will describe the data flow between the various processes in the system.

**2.7.1 DFD Diagrams**

Context Diagram (DFD0):

The system consists of 2 main processes: acquiring data to help recognize the DOM elements of the website and a script composition/execution process



diagram - context diagram

Element Mapping:

The user selects elements in the browser control The elements are then saved in the DB.



diagram – element mapping

Edit element map:

The user (script programmer) selects and marks the priority of the DOM element attribute according to which the element will be identified in the script.



diagram - edit element map

Edit Script:

In this process script is being composed and stored in the DB



diagram – edit script

Script run:

DOM element mapped to site objects are being monitored and manipulated according to the script commands (steps)



diagram – script run

**3 Specific Requirements**

**3.1 Functional requirements.**

* The system shall save all its data in a DB based on MS SQL sdf format (see diagram 6):
  + Element mapping between application objects and DOM elements. This will be done utilizing the website/element/recognition tables where a match of several site object properties will be used to select the corresponding DOM element
  + Script elements that compose scripts, which consist of script/step/condition/action tables

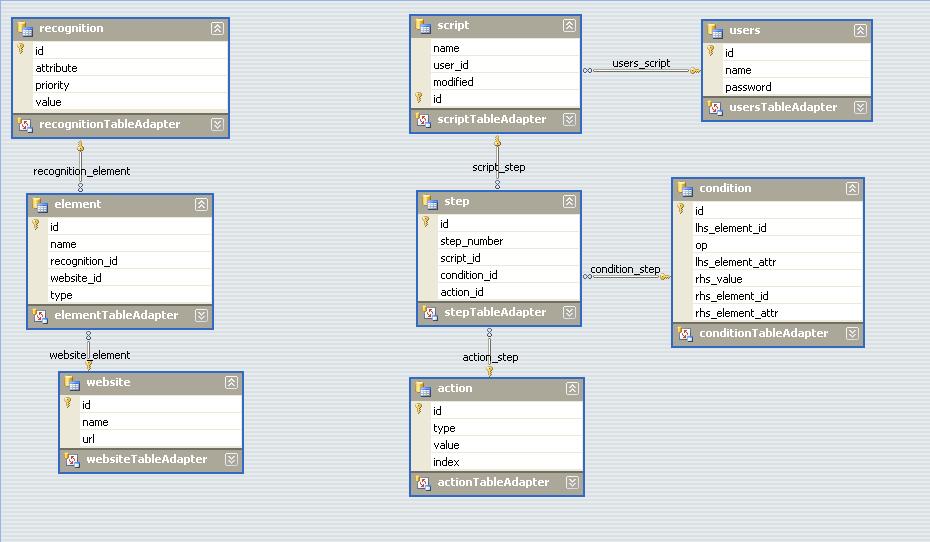


Diagram - ERD

* The application shall have 4 main forms:
  + Browser control form via which the user will select the relevant elements
  + Map editing form where the user will tweak the table that will help in recognizing DOM elements. The table will map between a site object and DOM element attributes.
  + Script scheduling form where the user will determine the time and frequency of script execution.
  + Script execution status and log.
* Element selection will be done using mouse hover over the element and pressing the middle mouse button for confirmation. When the mouse hovers over an element it shall be surrounded by a red rectangle.
* Script execution shall reuse the browser control form.
* Log of execution progress shall be displayed and written to a file.
* Workflow shall be implemented according to the DFD diagrams in section 2.7.1

**3.2 Performance requirements.**

* The system will support at least 3 simultaneous website interaction.
* Element interaction (click/text fill) shall be done in 1 second max.

**3.3 Interface requirements.**

NA

**3.4 Resource requirements.**

* X86 based cpu
* MS windows XP or greater
* 2 GB memory
* 20MB hard disk space

**3.5 Verification requirements.** As in test plan document

**3.6 Security requirements.**

All site and user data shall be stored in a password protected DB file.

# נספח ב2 - (Software Test Plan (STP

**(מבוסס על IEEE Std 829-1998)**

1. **Introduction**

This test plan will verify the entire requirements in the Auto Web Agent SRD were implemented correctly. The philosophy of testing is to test each component as isolated as possible (Unit test each component by building a test class for each component class) and then perform tests on the system as a whole.

Three synthetic websites (i.e. html files) with assorted types of elements will be used as the test subjects.

**Test items**

* + - Element processing
    - Script processing

1. **Features to be tested**

* Test functionality of element recognition:
  + Select text input element and verify it is entered in the DB correctly.
  + Select radio button and verify it is entered in the DB correctly.
  + Select a list element and verify it is entered in the DB correctly.
  + Select button and verify it is entered in the DB correctly.
  + Select checkbox and verify it is entered in the DB correctly.
* Test functionality of map editor
  + Change name of several elements and verify they are changed in the DB
  + Change the recognition priorities/attributes of an element and verify it is reflected in the DB
* Test script editor
  + Create new script
  + Add several steps
  + Delete several steps
  + Save script to DB.
* Test script scheduling and execution
  + Schedule a script to run after 10 minutes, check it executes correctly.
  + Schedule a script to run after 1 hour for 3 times, check it executes correctly.
  + Schedule a script to loop indefinitely, check it executes correctly.

1. **Features not to be tested:**
   1. the open source WatiN library.

1. **Environmental needs –** as in project proposal section 4.
2. **Schedule** – as in Gantt chart.

# נספח ב3 - (Software Project Management Plan (SPMP

**(מבוסס על IEEE Std 1058)**

1. **Introduction** - Project Management Plan Overview

This appendix will describe the progress made and the progress planed of the project execution.

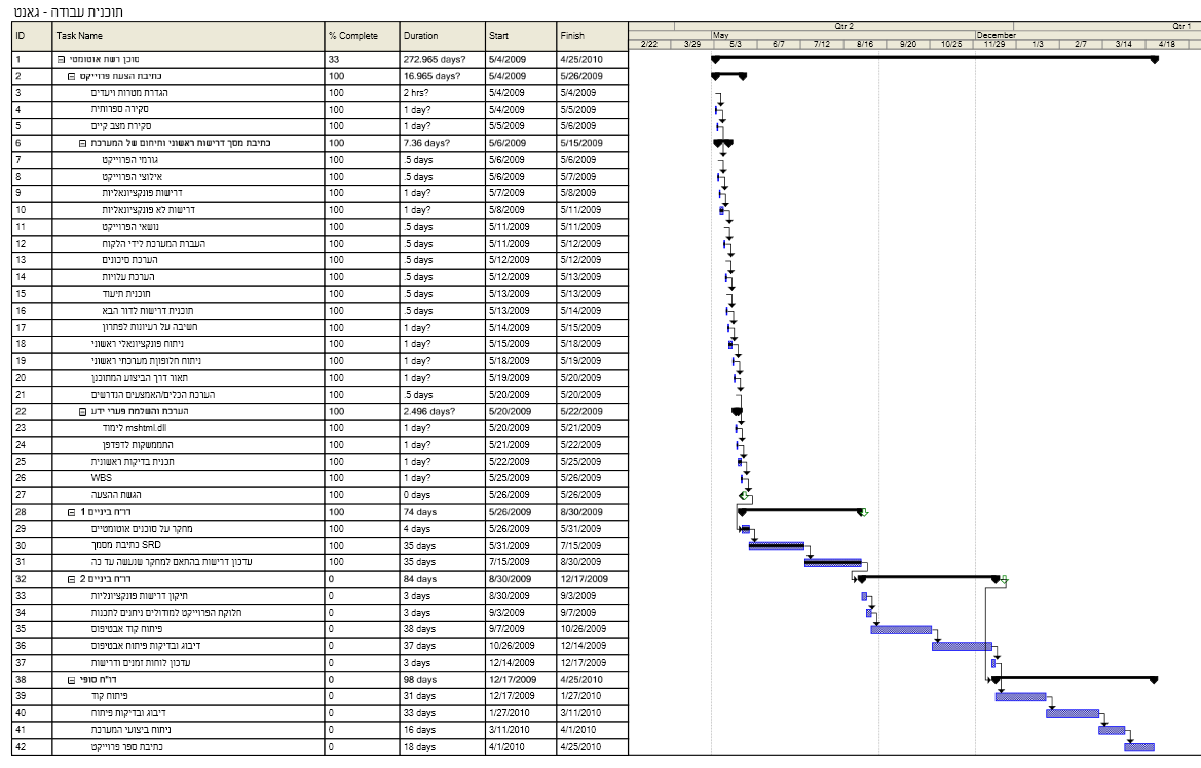
1. **Risk Management** - Describes updates of risks and managing them.

No new risk was identified (all risks are listen in the project proposal appendix)

1. **Evaluation of the SPMP**

The progress of the project is more or less on track except for a minor delay in deliver the first report. Technology selection and feasibility have been made in the form of mini-projects for proof of concept purposes. The next thing is to begin coding the prototype module by module and then integrate it into a solid and robust application.

1. **Schedule –** Updated and Detailed Work Plan.



נספח 1 – הצעת הפרויקט:

****

**מחלקת הנדסת תוכנה**

שם הפרויקט: סוכן רשת אוטומטי

Project Name: Automatic Web Agent

הצעת פרויקט

|  |  |
| --- | --- |
| שם הסטודנט: | קובי הרשקוביץ |
| מספר תעודת זהות: | 025632654 |
| שם המנחה: | אלעד הוגן |
| תאריך ההגשה: | 25/05/09 |

**תוכן ענינים**

[מבוא 4](#_Toc229036130)

[מטרות, יעדים ומדדים 5](#_Toc229036131)

[סקירה ספרותית 6](#_Toc229036132)

[סקירת מצב קיים 7](#_Toc229036133)

[מסמך דרישות ראשוני ותיחום של המערכת 7](#_Toc229036134)

[גורמי הפרויקט 7](#_Toc229036135)

[תכלית הפרויקט 7](#_Toc229036136)

[לקוחות הפרויקט 7](#_Toc229036137)

[משתמשי הפרויקט 7](#_Toc229036138)

[אילוצי הפרויקט 7](#_Toc229036139)

[אילוצים ראשיים/קריטיים 7](#_Toc229036140)

[הגדרות שמות, קיצורים וראשי תיבות 7](#_Toc229036141)

[עובדות והנחות הקשורות לפרויקט 8](#_Toc229036142)

[דרישות פונקציונאליות 8](#_Toc229036143)

[תחום העבודה 8](#_Toc229036144)

[תחום המוצר 8](#_Toc229036145)

[דרישות מידע ופונקציונאליות 8](#_Toc229036146)

[דרישות לא פונקציונאליות 8](#_Toc229036147)

[דרישות עיצוב 8](#_Toc229036148)

[דרישות שימוש וממשק משתמש 8](#_Toc229036149)

[דרישות ביצועים 8](#_Toc229036150)

[דרישות הפעלה 8](#_Toc229036151)

[דרישות תחזוקה ותמיכה 9](#_Toc229036152)

[דרישות אבטחה 9](#_Toc229036153)

[דרישות תרבותיות ופוליטיות 9](#_Toc229036154)

[דרישות חוקיות 9](#_Toc229036155)

[נושאי הפרוייקט 9](#_Toc229036156)

[נושאים פתוחים 9](#_Toc229036157)

[פתרונות מדף 9](#_Toc229036158)

[בעיות חדשות 9](#_Toc229036159)

[משימות 9](#_Toc229036160)

[העברת המערכת לידי/לניהול הלקוח 9](#_Toc229036161)

[סיכונים 9](#_Toc229036162)

[עלויות 9](#_Toc229036163)

[תיעוד ולימוד משתמשים 10](#_Toc229036164)

[דרישות לדור הבא 10](#_Toc229036165)

[רעיונות לפתרון 10](#_Toc229036166)

[ניתוח פונקציונאלי ראשוני 10](#_Toc229036167)

[ניתוח חלופות מערכתי ראשוני 13](#_Toc229036168)

[תאור דרך הביצוע המתוכנן 13](#_Toc229036169)

[האמצעים/הכלים הנדרשים 13](#_Toc229036170)

[פערי ידע שעל הסטודנט להשלים 14](#_Toc229036171)

[תוצרי הפרויקט 14](#_Toc229036172)

[תכנית בדיקות ראשונית 15](#_Toc229036173)

[ניהול סיכונים 15](#_Toc229036174)

[תוכנית ניהול הפרויקט 16](#_Toc229036175)

[Work Breakdown Structure 16](#_Toc229036176)

[תוכנית עבודה של הפרויקט 17](#_Toc229036177)

[רשימת מקורות: 17](#_Toc229036178)

# מבוא

כיום השימוש באינטרנט נעשה רווח כמעט בכל בית ובאמצעותו ניתן לבצע משימות רבות ומגוונות החל מקבלת מידע בידורי וכלה בפעולות בנקאיות. במרבית מן הפעולות נדרש המשתמש להזדהות, לבצע מעקב אחר נתונים באתר, ולבצע פעולות על סמך ידע שלמד מאתרים אחרים. פעולות אלו במרבית הזמן זהות/דומות בין האתרים וכמו כן ביצוען עלול לדרוש תשומת לב וזמן רב מן המשתמש.

הפרוייקט המוצע הינו סוכן רשת אוטומטי הבא להחליף מטלות שגרתיות וגוזלות זמן של משתמש דפדפן ממוצע/מתקדם. באמצעות המערכת המוצעת יוכל המשתמש לתכנן תסריט של פעולות מסויימות באתרים מסויימים בהתאם למידע הנדגם מאתרים אחרים ולקבוע את זמן ביצוע התסריט.

לדוגמא ניתן לקבוע תסריט שבו נדגם כל 4 שעות אתר סקי שהמשתמש מחבב במיוחד (לדוגמא www.champery.com) , התסריט מכוון לזהות את השדה של גובה השלד באתר וברגע שגובה השלג מגיע ל-110cm התסריט שולף מאתר ה-"דקה ה-90" את מחירי חבילות הנופש לאתר ושולח דוח מפורט עם הנתונים באי-מייל למשתמש.

דוגמא נוספת יכולה להיות תסריט שיאריך את זמן השאלת ספר באתר הספריה של מכללת אפקה ,שאינו מצוייד כיום בסוכן אוטומטי להארכת השאלות, פעם בשלושה ימים.

יחודיות הפרוייקט היא שבניגוד לסוכנים אוטומטיים הנמצאים באתר מסויים ונשלטים מתוך האתר הפרוייקט הנ"ל מציע סוכן שיתוכנת בצד הלקוח ולא בצד השרת, דבר אשר יאפשר לו נוחויות ושליטה מירבית להשגת מטרותיו.

# מטרות, יעדים ומדדים

**מטרות:**

רבות מהפעולות הפיננסיות, הלימודיות והבידוריות נעשות באמצעות דפדפן ברשת האינטרנט:

* רישום לאתרים
* קניה/מכירה של מניות
* הזמנת כרטיסים
* הגשת מטלות
* בחירת קורסים

רבות ממטלות אלו הן מחזוריות, שגרתיות ועלולות להתיש את המשתמש ו/או לגרום לו לבצע שגיאות באופן הכנסת הנתונים.

מטרת הפרוייקט הנגזרת מבעיה זו היא בניית אפליקציה אשר תבצע מטלות שגרתיות וצפויות מראש בצורה אוטומטית במקום משתמש דפדפן אנושי

**יעדים:**

ניתן לפרוט מטרת על זו ל-3 יעדים:

* ניטור אתר בצורה מחזורית
* ביצוע פעולות אוטומטיות בדפדפן
* התראה למשתמש ע"פ בקשתו על אירוע באתר מסויים.

**מדדים:**

המדדים ע"פ תימדד הצלחת הפרוייקט הם:

* אחוז האתרים הנתמכים מתוך עשרת האתרים הראשונים בחיפוש בגוגל יעמוד על 80% בקטגוריות פיננסים, בידור, ספורט (דוגמא לחיפוש יכולה להיות המילה 'פיננסי')
* נוהל כניסה לאתר הספרייה במכללה והארכת השאלה אם קיימים ספרים מושאלים יסתיים בהצלחה.
* נוהל דגימת אתר לשינויים בתדירות של 6 פעמים בדקה יסתיים בהצלחה (ידווח בצורה נכונה על שינוי).
* משתמש חדש יוכל להגדיר תסריט אוטומטי של דגימת 2 אתרים ואינטראקציה ביניהם בתוך כשעה.

# סקירה ספרותית

סוכני רשת הם מערכות תוכנה מורכבות אשר פועלות ברשת האינטרנט העולמית (www), ברשת הארגונית או ברשתות מקומיות. סוכני רשת מתוכננים לבצע מגוון משימות החל מאגירה וניתוב של מידע וכלה בחיפוש קיטלוג וסינון מידע.

הספרות מתארת מסגרת לתכנון וניתוח של סוכני רשת ומערכות סוכנים בהתבסס על מודלים מתמטיים של הסביבה בה הם פועלים. יש 3 אבני בניין מרכזיות לבניית סוכני רשת:

* מודלים סטוקאסטיים של גרף הרשת אשר משקללים את הפילוג והקישוריות של דפי רשת ונותנים קוים מנחים מרכזיים בתכנון סוכנים, ע"י הפקת תכונות ייחודיות של הסביבה בה הסוכנים פועלים.

הרשת יכולה להיראות כגרף מכוון אשר הצמתים בו הם דפי HTML סטטיים והקשתות בו הם קישורים מאתר אחד למשנהו. מאמר [ ‎18] מציע ומנתח מודלי גרפיים אקראיים שנוצרו ע"י תצפיות נסיוניות ברשת. המודלים שונים מהמודלים המקובלים בכך ש קשתות שנבחרו בצורה בלתי תלויה לא משתקפות בסטטיסטיקה שנצפתה ברשת, כך שקשתות במודלים אלו הינם תלויות סטטיסטית ביניהן. כמו כן נוספים צמתים לגרף כתלות בזמן. דבר זה משקף את העובדה שהרשת משתנה עם הזמן.

* שיטות למידה סטטיסטית המאפשרות לסוכנים ללמוד על סביבתם ע"י הפקת מודלים סטוקאסטיים של דף רשת ומבנה קישורים מקומיים. מאמר [ ] מתאר את הרשת כעולם שבו המוני קישורים מחברים בין מסמכי ודפי אתרים כך שנוצר גרף בעל סיבוכיות גבוהה וללא תקדים וזהו גרף הרשת. מאמר זה מציג גישה חדשה ללימוד מודלים הסתברותיים שך הרשת אשר יכולים לשמש כחלק בתהליך חיזוי קישוריות ומידע של מסמכי רשת. השיטה המתוארת במאמר זה היא שיטה הסתברותית לצמצום מימדים (probabilistic dimension reduction) אשר גוזרת ומאחדת את שיטת הניתוח הסמנטי המאוחר (Latent Semantic Analysis) ואת אלגוריתם קליינברג (Hubs-and-Authorities) בסביבה סטטיסטית.

דבר זה אמור להיות הצעד הראשון לפיתוח תשתית סטטיסטית של טכנולוגיות מידע ברשת. למרות שמאמר זה לא מתמקד ביישום מסויים, מגוון אלגוריתמים הפועלים בסביבת רשת יכולים לנצל את היתרונות בטכניקות המוצגות למטרות שונות כגון מנועי חיפוש, "זוחלי רשת" (Web crawlers) ומערכות סוכני מידע.

* אלגוריתמים לתכנון אוטונומי וקבלת החלטות בסביבת רשת לאפשר לסוכנים להשיג מטרות ולהסתגל לסביבה משתנה [16]

ניקח לדוגמא בעיות קלאסיות בתחום בסיסי נתונים/איחזור מידע כגון חיפוש בטקסט, כריית מידע וסיווג. הקלט לבעיות מסוג זה הוא בדרך כלל אוסף של נתונים/מסמכים. הרשת עם המבנה הנוסף שלה כגרף, מאפשרת את השיפור של טכניקות קיימות עם טכניקות מעולם הגרפים. במאמר הודגם שיפור זה באמצעות פתרונות מבוססי גרפים לבעיות הבאות: חיפש נושא, מיספור נושא, סיווג וזחילה (crawling).

# סקירת מצב קיים

כיום כאשר משתמש רוצה לבצע מעקב/שינוי נתונים באתר מסויים פרושות לפניו מספר אפשרויות:

תמיכה של האתר עצמו בסוכן רשת אשר מיידע את המשתמש לגבי נתון ייחודי לאתר אליו הוא שייך. לדוגמא האתר ebay מציע סוכן שיציע במכרז מחיר עבור פריט מסויים עד תקרה שיקבע המשתמש.

מעקב מחזורי אישי של המשתמש אחר שינויים באתר.

שימוש בתוכנות אוטומציה יקרות כדוגמת QFT או RFT.

QTP(Quick Test Pro ) היא חבילת תוכנה מבית מרקורי שבה ניתן לבצע אוטומציה לאפליקציות חלונאיות ורשתיות כגון אוטומציה לכתיבת דואר אלקטרוני ושליחתו או חקירת דף רשת והכנסת נתונים אליו.

RFT (Rational Functional Tester) היא גם חבילת תוכנה לבדיקות אוטומטיות מבית IBM המאפשרת גם לבצע בדיקות על אלמנטים מסוג Flex וטרמינלים כגון VT102.

מחירם של כל אחת מחיבולות אלו מגיע לכמה אלפי דולרים עבור רשיון למשתמש אחד.

# מסמך דרישות ראשוני ותיחום של המערכת

## ***גורמי הפרויקט***

### **תכלית הפרויקט**

להוות כלי למעקב, והתראה על שינויים באתרי אינטרנט וביצוע פעולות מתוכנתות מראש כתגובה למידע הנכרה מהם, באותם אתרים עצמם או אתרי מטרה אחרים.

### **לקוחות הפרויקט**

משתמשי דפדפן מתקדמים הרוצים לבצע פעולות אוטומטיות בדפדפן ללא מגע יד אדם.

### **משתמשי הפרויקט**

ראה לקוחות

## ***אילוצי הפרויקט***

### **אילוצים ראשיים/קריטיים**

תמיכה ב-IE בלבד.

### **הגדרות שמות, קיצורים וראשי תיבות**

|  |  |  |
| --- | --- | --- |
| שם/קיצור/ר"ת | פירוש באנגלית | פירוש בעברית |
| IE | Internet Explorer | דפדפן מבית מיקרוסופט |
| HTML | Hyper Text Markup Language | שפה לתיאור טקסקט עשיר |
| DOM | Document Object Model | מודל אובייקט מסמך |
| אלמנט | Element | פריט בדף HTML |
| URL | Uniform Resource Locator | מאתר משאבים אחיד |
| Hostname | Server name part of the URL | שם השרת ב-URL |
| WWW | World Wide Web | רשת האינטרנט העולמית |

### **עובדות והנחות הקשורות לפרויקט**

הדף הנבדק לא יכיל אובייקטי Flash ו-Java

למשתמש יש ידע בסיסי ב-HTML

הדפדפן יהיה מסוג IE.

## ***דרישות פונקציונאליות***

### **תחום העבודה**

רשת האינטרנט (WWW)/אינטרהנט.

### **תחום המוצר**

ניטור ואוטומציה של אתרים.

### **דרישות מידע ופונקציונאליות**

* בסיס נתונים לשמירת מיפוי אלמנטים בדף לאובייקטים, ולשמירת טבלת מצב-פעולה (trigger-action table)
* ממשק משתמש:
  + מיפוי אלמנטים בדף לאובייקטים בבסיס נתונים.
  + כתיבת תסריט פעולה.
  + קביעת תזמון ריצה.
* יכולת זיהוי אלמנטים בדף HTML.
* יכולת הכנסת נתונים אוטומטית בדף HTML.
* יכולת דגימה מחזורית של שדות מסויימים באתרים מסויימים
* יכולת התראה על אירוע למשתמש במצעות חלון /או דואר אלקטרוני
* המשתמש יוכל להגדיר באילו תנאים תבוצע פעולה, אותה יוכל להגדיר גם כן (לדוגמה לחיצה על כפתורים ומילוי שדות כאשר ערך של שדה מסויים ישתנה)

## ***דרישות לא פונקציונאליות***

### **דרישות עיצוב**

הממשק יעוצב כך שמשתמש דפדפן ממוצע יוכל להבין בצורה אינטואיטיבית את פעולתו.

### **דרישות שימוש וממשק משתמש**

הממשק יהיה מעוצב כממשק לשוניות כאשר:

* הלשונית הראשונה תכיל את ממשק מיפוי האלמנטים
* הלשונית השניה תכיל את ממשק כתיבת התסריטים
* הלשונית השלישית תכיל את ממשק תזמון והרצת התסריטים.

### **דרישות ביצועים**

המערכת תפעל במהירות גבוהה ממהירות עבודה של משמתמש אנושי.

### **דרישות הפעלה**

* חיבור לרשת.
* IE מותקן.

### **דרישות תחזוקה ותמיכה**

תיקוני באגים ובוצעו ע"פ דרישות הלקוח.

### **דרישות אבטחה**

סיסמאות ופרטים אישיים יוצפנו בבסים הנתונים

### **דרישות תרבותיות ופוליטיות**

NA

### **דרישות חוקיות**

NA

## ***נושאי הפרוייקט***

### **נושאים פתוחים**

* האם יוכלו לרוץ כמה תסריטים במקביל
* האם להכניס מנגנון שיתמוך בשינויים באתר/ים

### **פתרונות מדף**

* Newbie (<http://www.newbielabs.com/webautomation.html>)
* iMacros (<http://wareseeker.com/publisher/imacros-website-testing-and-we/38865/>(
* iRobot (<http://www.irobotsoft.com>)
* Selenium (<http://seleniumhq.org/>)
* Rational Functional Tester (<http://www.ibm.com/software/awdtools/tester/functional/index.html>)
* QTP (<http://mercuryquicktestprofessional.blogspot.com/>)

### **בעיות חדשות**

NA

### **משימות**

המשימות הנדרשות מתוארות ב-WBS ובתוכנית העבודה.

## ***העברת המערכת לידי/לניהול הלקוח***

מוצר מדף – יתבצע באמצעות תיעוד ודוגמאות שימוש.

## ***סיכונים***

ראה ‏13 (ניהול סיכונים)

## ***עלויות***

פרוייקט תוכנה בלבד – ההשקעה היא זמן פיתוח בלב

## ***תיעוד ולימוד משתמשים***

התיעוד לפרוייקט יהיה בהערות בגוף הקוד, בדוחות הביניים ובספר הפרוייקט.

## ***דרישות לדור הבא***

לא ידוע עדיין.

## ***רעיונות לפתרון***

ראה ניתוח פונקציונלי ראשוני לפירוט הרעיון

# ניתוח פונקציונאלי ראשוני

פעולת המערכת נחלקת לשלושה שלבים:

* איסוף/מיפוי האלמנטים ה-"מעניינים"

בשלב זה המשתמש גולש לאתר המטרה באמצעות הדפדפן המוטמע באפליקציה, ומסמן אלמנטים רלוונטיים בדף באמצעות העכבר. בלחיצה על מקש במקלדת ייטען השדה המסומן לטופס מיפוי האלמנטים באפליקציה. טופס זה מכיל מיפוי בין האלמנט ב-DOM לבין הייצוג הפנימי שלו במערכת אשר יישמש ביצירת התסריטים, וכן מידע נוסף על האלמנט כגון הטיפוס שלו, הערך הנוכחי שלו וכו'.

מיפוי המחדל עבור שם אובייקט יהיה ה-id של האלמנט, אך המשתמש יוכל לשנות זאת בטופס זה.

* הגדרת תסריט פעולה

בשלב זה המשתמש מגדיר תסריט פעולה המורכב מזוגות סדורים של תנאי קיום ופעולה (להלן יקרא זוג כזה "כלל"). תנאי הקיום יכול לבדוק ערך של אלמנט (המיוצג ע"י אובייקט) ואם הוא עומד בתנאי אז יתבצע חלק הפעולה של הכלל.

לאחר הגדרת כל כלל יתבצע מעבר בדיקה לוודא את תקינות הכללים בתסריט.

* תזמון/הרצה של תסריטים

בשלב זה בוחר המשתמש תסריט להרצה וניתנת לו האפשרות לקבוע מתי ירוץ התסריט ובאיזו מחזוריות.

מפת האלמנטים וכן התסריטים ישמרו בבסיס נתונים לצורך המשכיות (Persistence).

## ***תרשימי DFD***

תרשים הקשר:



תרשים ‏6‑1 Context Diagram(DFD0)

מיפוי אלמנטים:



תרשים ‏6‑2 מיפוי אלמנטים

עריכת מפת אלמנטים:



תרשים ‏6‑3 עריכת מפת אלמנטים

עריכת תסריט:



תרשים ‏6‑4 עריכת תסריט

הרצת תסריט:



תרשים ‏6‑5 הרצת תסריט

# ניתוח חלופות מערכתי ראשוני

|  |  |  |  |
| --- | --- | --- | --- |
| נושא לחלופה | QTP | iMacros | RFT |
| מחיר($) | 35000-70000 | 50-500 | 6500-9000 |
| כיסוי פונקצינאלי | מושלם | חסרה שליטה בתזמון | מושלם |
| קלות שימוש | מסובך – יש ללמוד VBScript | פשוט | בינוני – יש לדעת JAVA ברמה בינונית/נמוכה. |
| Scripting | VBScript | Proprietary | JAVA |

# תאור דרך הביצוע המתוכנן

* הגדרת סכמה לבסיס הנתונים – יש להגדיר את מבנה הטבלאות כגון מבנה טבלת המיפוי ומבנה טבלאות התסריטים
* הגדרת שפת תסריט – יש להחליט על האופן בו המשתמש יתכנת תסריטים במערכת. האפשרויות הנשקלות הן שימוש באבני בניין קבועות ליצירת צעדי תסריט כך שכל צעד יורכב מבחירת מספר אלמנטים קבוע באמצעות פקדים חלונאיים. אפשרות נוספת היא כתיבת תסריט בשפת C# וניתוחו באמצעות מודול הידור תסריטים.
* חלוקת המערכת למודולים – רכיבי המערכת יחולקו למודולים הניתנים להכלה במחלקות.
* הגדרת המחלקות הנדרשות בכל מודול ומודול – כל מחלקה תכיל פונקציה לוגית בודדת.
* קידוד ובדיקת כל מודול בזה אחר זה- הקידוד יעשה בשפת C# ובדיקתו תיעשה ב-NUNIT.
* אינטגרציה של כל המודולים – בשלב זה תיבנה בפעם הראשונה כל האפליקציה כמערכת אחת.
* בדיקות – ביצוע בדיקות המערכת ווידוא שכל המטרות הושגו. כמו כן יזוהו כל הבאגים ויתועדו עבור שלב התיקון.
* תיקוני באגים ובדיקות חוזרות – ניפוי השגיאות עד לפעולה מיטבית של המערכת.
* תיעוד – כתיבת הדוחות וספר הפרוייקט. כמו כן התיעוד בתוך הקוד יעשה במהלך כתיבת הקוד.

# האמצעים/הכלים הנדרשים

**סביבת פיתוח:**

סביבת פיתוח VS2008.

דפדפן IE8

חיבור לרשת

**סביבת ריצה:**

חיבור לרשת

דפדפן IE8

.NET runtime framework 3.5

# פערי ידע שעל הסטודנט להשלים

לימוד הממשקים הנדרשים (כגון mshtml.dll, docview.dll, DOM) בעיקר מאתר MSDN

# תוצרי הפרויקט

**דוח ביניים** **1** – יכיל את כל אשר נדרש ע"פ ההגדרה במסמך "מבנה תוצרים – הנדסת תוכנה"

* + אבטיפוס.
  + תקציר
  + סקירה ספרותית
  + מטרות העבודה
  + סקר מצב קיים
  + ניתוח חלופות מערכתי
  + ניתוח חלופות טכנולוגיות
  + תיאור פיתרון נבחר
  + תוכנית בדיקות תוכנה
  + לוחות זמנים
  + ניהול סיכונים
  + רשימת מקורות.
  + מסמכים מצורפים SRD ,STP , SPMP. הצעת פרויקט.

**דוח ביניים 2** - יכיל את כל אשר נדרש ע"פ ההגדרה במסמך "מבנה תוצרים – הנדסת תוכנה"

* + תקציר
  + מטרות העבודה
  + סקירה ספרותית
  + אלטרנטיבות לתיכון המערכת
  + הפתרון שנבחר לתיכון המערכת
  + תוכנית בדיקות תוכנה
  + לוחות זמנים מעודכנים להמשך העבודה
  + הערכת ביצוע המשימות
  + ניהול סיכונים
  + רשימת מקורות
  + מסמכים: SRD, SDD, STD, SPMP. הצעת פרויקט.
  + אפליקציה מושלמת (קובץ התקנת אפליקציה, חוברת הפעלה).

**ספר פרוייקט** - יכיל את כל אשר נדרש ע"פ ההגדרה במסמך "מבנה תוצרים – הנדסת תוכנה"

* + יסכם את דו"חות הביניים
  + ישלב דרישות מערכת ואפיון מערכת
  + יפנה למסמכים שכתבנו בדוחות הביניים
  + ארכיטקטורת המערכת
  + דווח בדיקות תוכנה
  + דוגמאות להפעלת המערכת
  + יתאר ביצועים עלות ואמינות
  + יתאר את הצלחת המערכת או אי ההצלחה.
  + המסמכים שנעשו במשך העבודה על הפרויקט.

**מערכת**

* + קבצי התקנה
  + קבצי עזרה והדרכה למשתמש
  + אבטיפוס ראשוני
  + מצגת פרויקט
  + מערכת סופית

# תכנית בדיקות ראשונית

* בניית דפי דמה המכילים מגוון רב של סוגי אלמנטים וניסוי המערכת עליהם.
* בדיקת אוטומציה של חיפוש באתר הבית של גוגל.
* בדיקת ניטור שדה באתר ושליחת מייל כאשר שדה זה עובר ערך מסויים
* בדיקת מנוע יצירת תסריטים

# ניהול סיכונים

|  |  |  |  |
| --- | --- | --- | --- |
| סיכון | סיכוי | השפעה | דרך מניעה |
| חריגה מלוחות זמנים | 50% | אי הגעה לאבני דרך בפרוייקט | ניסיון להקדים ביצוע של משימות והקדשת זמן נוסף לפרוייקט |
| כשל חומרתי בסביבת הפיתוח/ריצה | 20% | עיכוב של עד כשבועיים | גיבוי כל איטרציה במערכת ניהול קוד |
| אי יכולת לזהות אלמנטים מיוחדים בדף | 50% | אי יכולת להשתמש באלמנטים אלו – ירידה בפונקציונאליות המובטחת | עדכון סעיף האילוצים בשלב מוקדם ככל האפשר |
| דפים דינמיים (Ajax, DHTML) | 80% | שיבוש יכולת זיהוי האלמנטים | התקנת מנגנוני המתנה לאלמנטים לא קיימים. |

# תוכנית ניהול הפרויקט

## ***Work Breakdown Structure***

## ***תוכנית עבודה של הפרויקט***



תרשים ‏14‑1

# רשימת מקורות:

1. Jon Kleinberg. **Authoritative sources in a hyperlinked environment**. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997. [[pdf]](http://www.cs.cornell.edu/home/kleinber/auth.pdf)
2. Soumen Chakrabarti, Martin van den Berg, Byron Dom, **Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery**, WWW8 [[html]](http://www8.org/w8-papers/5a-search-query/crawling/index.html)
3. Steve Lawrence and Lee Giles, **Searching the World Wide Web,** Science 1998 [[pdf]](http://www.neci.nj.nec.com/~lawrence/papers/search-science98/search-science98.pdf)
4. Jeffrey Dean, Monika R. Henzinger, **Finding Related Pages in the World Wide Web**, WWW8 [[html]](http://www8.org/w8-papers/4a-search-mining/finding/finding.html)
5. Oren Zamir and Oren Etzioni, **Grouper: A Dynamic Clustering Interface to Web Search Results**, WWW8, [[html]](http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html)
6. Rajan M. Lukose and Bernardo A. Huberman, **Surfing as a Real Option** [[abstract]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/options.html)
7. Bernardo A. Huberman and Rajan M. Lukose, **Social Dilemmas and Internet Congestion** [[abstract]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/storms.htm)
8. Bernardo A. Huberman, Peter L.T. Pirolli, James E. Pitkow, and Rajan M. Lukose, **Strong Regularities in World Wide Web Surfing**, Nature ??? [[abstact]](http://www.parc.xerox.com/spl/groups/dynamics/abstracts/InternetEcology/surfing.html)
9. William W. Cohen, Wei Fan, **Web-Collaborative Filtering: Recommending Music By Spidering the Web**, WWW9 [[html]](http://www9.org/w9cdrom/266/266.html)
10. Mike Perkowitz Oren Etzioni, **Towards Adaptive Web Sites: Conceptual Framework and Case Study**, WWW8 [[html]](http://www8.org/w8-papers/2b-customizing/towards/towards.html)
11. Marc Langheinrich, Atsuyoshi Nakamura, Naoki Abe, Tomonari Kamba, Yoshiyuki Koseki, **Unintrusive Customization Techniques for Web Advertising**, WWW8 [[html]](http://www8.org/w8-papers/2b-customizing/unintrusive/unintrusive.html)
12. Jim Hendler, **Is there an intelligent agent in your future?** Nature, Web matters, 11 March 1999. [[html]](http://helix.nature.com/webmatters/agents/agents.html)
13. Steve Kirsch, **The future of Internet search (keynote address)**, SIGIR 1999 [[pdf]](http://www.acm.org/pubs/articles/proceedings/ir/290941/p1-kirsch/p1-kirsch.pdf)
14. G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure, Proceedings of the 8th International Computing and Combinatorics Conference (COCOON), 2002.
15. [G. Pandurangan](http://www.cs.brown.edu/people/gopal) , [P. Raghavan](http://theory.stanford.edu/people/raghavan/index.html) , and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**Building Low-Diameter P2P Networks**](http://www.cs.brown.edu/research/webagent/focs-2001.pdf). Proceedings of the 42th IEEE Symp. on Foundations of Computer Science. 2001.
16. S.R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**The Web as a graph**](http://www.cs.brown.edu/research/webagent/pods-2000.pdf). Proceedings of the 19th ACM Symposium on Principles of Database Systems, pp 1-10, 2000.
17. Thomas Hofmann, [**Learning Probabilistic Models of the Web**](http://www.cs.brown.edu/research/webagent/Hofmann-SIGIR00.pdf), ACM SIGIR 2000
18. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and [Eli Upfal](http://www.cs.brown.edu/people/eli), [**Stochastic models for the Web graph**](http://www.cs.brown.edu/research/webagent/focs-2000.pdf). Proceedings of the 41th IEEE Symp. on Foundations of Computer Science. 2000.

1. WatiN (pronounced "what-in") aims to bring an easy way to automate Web Application Testing in .Net. Inspired by Watir, WatiN has grown into a feature rich and stable framework, offering automation of Internet Explorer and Firefox. [↑](#footnote-ref-2)