



NEXT PAGE PREDICTION WITH POPULARITY BASED PAGE RANK, DURATION  
BASED PAGE RANK AND SEMANTIC TAGGING APPROACH

A THESIS SUBMITTED TO  
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES  
OF  
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BANU DENİZ YANIK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR  
THE DEGREE OF MASTER OF SCIENCE  
IN  
COMPUTER ENGINEERING

FEBRUARY 2012

Approval of the thesis:

**NEXT PAGE PREDICTION WITH POPULARITY BASED PAGE RANK, DURATION  
BASED PAGE RANK AND SEMANTIC TAGGING APPROACH**

submitted by **BANU DENİZ YANIK** in partial fulfillment of the requirements for the degree  
of **Master of Science in Computer Engineering Department, Middle East Technical Uni-  
versity** by,

Prof. Dr. Canan Özgen  
Dean, Graduate School of **Natural and Applied Sciences**

---

Prof. Dr. Adnan Yazıcı  
Head of Department, **Computer Engineering**

---

Assoc. Prof. Dr. Pınar Şenkul  
Supervisor, **Computer Engineering Department, METU**

---

**Examining Committee Members:**

Prof. Dr. Nihan Kesim Çiçekli  
Computer Engineering Dept., METU

---

Assoc. Prof. Pınar Şenkul  
Computer Engineering Dept., METU

---

Prof. Dr. İsmail Hakkı Toroslu  
Computer Engineering Dept., METU

---

Dr. Ayşenur Birtürk  
Computer Engineering Dept., METU

---

Assist. Prof. Osman Abul  
Computer Engineering Dept., TOBB Uni.

---

**Date:**

---

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last Name: BANU DENİZ YANIK

Signature :



# ABSTRACT

## NEXT PAGE PREDICTION WITH POPULARITY BASED PAGE RANK, DURATION BASED PAGE RANK AND SEMANTIC TAGGING APPROACH

Yanık, Banu Deniz

M.Sc., Department of Computer Engineering

Supervisor : Assoc. Prof. Dr. Pınar Şenkul

February 2012, 154 pages

Using page rank and semantic information are frequently used techniques in next page prediction systems. In our work, we extend the use of Page Rank algorithm for next page prediction with several navigational attributes, which are size of the page, duration of the page visit and duration of transition (two page visits sequentially), frequency of page and transition. In our model we define *popularity* of transitions and pages by using duration information and use it in a relation with page size and visit frequency factors. By using the popularity value of pages, we bias conventional Page Rank algorithm and model a next page prediction system that produces page recommendations under given top-n value. Moreover we extract semantic terms from web URLs in order to tag pages semantically. The extracted terms are mapped into web URLs with different level of details in order to find semantically similar pages for next page recommendations. With this tagging, we model another next page prediction method which uses Semantic Tagging (ST) similarity and exploits PPR values as a supportive method. Moreover we model a Hybrid Page Rank (HPR) algorithm that uses both Semantic Tagging based approach and Popularity Based Page Rank values of pages together in order to investigate the effect of PPR and ST with equal weights. In addition, we investigate the effect of local (a synopsis of directed web graph) and global (whole directed web graph) modeling on next

page prediction accuracy.

Keywords: Next Page Prediction, Page Rank Algorithm, Semantic Tagging, Recommendation System

# ÖZ

## POPÜLERLİĞE GÖRE SAYFA SIRALAMASI, SAYFADA KALIŞ SÜRELERİNE GÖRE SAYFA SIRALAMASI VE SEMANTİK ETİKETLENMELERİNE GÖRE BİR SONRAKİ SAYFANIN ÖNGÖRÜLMESİ

Yanık, Banu Deniz

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Doç. Dr. Pınar Şenkul

Şubat 2012, 154 sayfa

Sayfa sıralaması ve semantik bilgi kullanımı bir sonraki sayfa öngörümünde sıkça tercih edilen bir yöntemdir. Çalışmamızda, bir sonraki sayfa öngörümünü Sayfa Sıralaması algoritmasıyla desteklerken bazı sayfa dolaşılmasına bağlı verilerle destekleyecek şekilde genişletilir. Bu veriler, sayfanın boyutu, sayfada kalma süresi, geçişin gerçekleşme süresi, sayfanın ve geçişin frekans değerleri olarak sıralanabilir. Modelimizde, sayfa ve geçiş *popülaritesi*, sayfa ve geçişe ait zaman bilgieri ve bu bilgilerin sayfa boyutuyla olan ilişkisi ve sayfa ve geçiş frekanslarıyla ilişkilendirilerek tanımlanmıştır. Popülerlik faktörü kullanılarak geleneksel Sayfa Sıralama algoritması yönlendirilerek, önerilen sayıda sayfayı öneren bir tavsiye sistemini gerçekleştirilir. Bunun yanında, web URL'lerinden semantik terimler çıkarılmıştır. Çıkarılan semantik terimler de seviyeli olarak web URL'ler ile etiketlenmiştir. Bu etiketleme sayesinde benzer etiketleme özelliğinden yararlanılarak semantik olarak benzer sayfaların analiz edilmesine imkan sağlanmıştır. Böylelikle bir sonraki sayfa öngörümünde Semantik Etiketleme olarak birbirine benzeyen sayfaların önerilmesi modellenmiştir. Bu modelleme kendi içinde destekleyici yöntem olarak Popüler Sayfa Sıralaması değerlerini kullanmaktadır. Ek olarak Melez Sayfa Sıralaması yöntemi ile de Semantik Etiketleme ve Popüler Sayfa

Sıralaması deęerleri eřit aęırlıkla da kullanılmıř ve bunun bir sonraki sayfa ngrmne etkisi arařtırılmıřtır. Ayrıca yerel (ynl web grafięinin sinopsisinin) ya da genel (ynl web grafięinin tamamının) modellemenin bir sonraki sayfa ngrmne etkisi de bu alıřma kapsamında arařtırılmıřtır.

Anahtar Kelimeler: Bir Sonraki Sayfanın ngrm, Sayfa Sıralama Algoritması, Semantik Etiketleme, Tavsiye Sistemi

*To my beloved father Adem YANIK, who never had a chance to learn that we are colleagues.*

## **ACKNOWLEDGMENTS**

First of all, I would like to thank my supervisor Pınar Şenkul for all her support and patience that she has given to me. Her friendly and positive behavior always gave me belief and courage from the beginning of this work.

I would also like to thank Çağlar Günel, my husband and my best friend. I could not complete this research without him. Furthermore I would like to thank my family for their support. They always believe in me and help me to survive.

I want to thank to my project manager and my colleagues for their support on me especially in statistical methodologies and other experiences on scientific researches and I would like to thank to all of my friends, who are so tolerant and understanding to me all the time.

# TABLE OF CONTENTS

ABSTRACT . . . . .	iv
ÖZ . . . . .	vi
ACKNOWLEDGMENTS . . . . .	ix
TABLE OF CONTENTS . . . . .	x
LIST OF TABLES . . . . .	xiii
LIST OF FIGURES . . . . .	xiv
CHAPTERS	
1 INTRODUCTION . . . . .	1
2 RELATED WORK . . . . .	6
2.1 Web Usage Mining with Markov Models . . . . .	6
2.2 Web Structure Mining and Page Rank Algorithm . . . . .	8
2.3 Web Content Mining and Semantic Web Mining . . . . .	9
3 BACKGROUND INFORMATION . . . . .	11
3.1 Markov Model and Directed Graph . . . . .	11
3.2 Conventional Page Rank Algorithm . . . . .	13
3.3 Usage Based Page Rank . . . . .	15
3.4 Semantic Annotation in Web . . . . .	16
4 DURATION AND POPULARITY BASED PAGE RANK . . . . .	17
4.1 Defining Sessions . . . . .	17
4.2 Duration Based Page Rank (DPR) . . . . .	18
4.3 Popularity Based Page Rank (PPR) . . . . .	19
4.4 PPR and DPR Calculations In Detail . . . . .	21
5 SEMANTIC TAGGING AND NEXT PAGE PREDICTION . . . . .	24
5.1 Semantic Tagging of URLs . . . . .	25

5.1.1	Captured Concepts in Experimental Setup . . . . .	27
5.1.2	Calculating Concept Similarity by an Example . . . . .	28
5.2	Next Page Prediction . . . . .	29
5.2.1	Next Page Prediction with DPR and PPR . . . . .	30
5.2.2	Next Page Prediction with Semantic Tagging Approach . .	31
5.2.3	Next Page Prediction with Hybrid Page Rank . . . . .	32
6	GENERAL ARCHITECTURE OF THE NEXT PAGE PREDICTION SYSTEM AND APPLICATION DOMAIN . . . . .	34
6.1	Data Set of the Application Domain . . . . .	34
6.2	Analyzing Server Logs . . . . .	36
6.2.1	Automatic Pruning of Web Server Logs . . . . .	36
6.2.2	Session Identification . . . . .	36
6.3	General Architecture of the Next Page Prediction System . . . . .	37
6.3.1	Page Rank Based System . . . . .	38
6.3.2	Semantic Tagging Based System . . . . .	39
6.3.3	Next Page Prediction System . . . . .	40
7	EXPERIMENTS AND EVALUATION . . . . .	42
7.1	3-Fold Cross Validation Experiments . . . . .	46
7.1.1	3-Fold Cross Validation with Top-2 Limits . . . . .	47
7.1.2	3-Fold Cross Validation with Top-4 Limits . . . . .	48
7.1.3	3-Fold Cross Validation with Top-8 Limits . . . . .	50
7.1.4	General Results . . . . .	53
7.2	5-Fold Cross Validation Experiments . . . . .	58
7.2.1	5-Fold Cross Validation with Top-2 Limits . . . . .	58
7.2.2	5-Fold Cross Validation with Top-4 Limits . . . . .	60
7.2.3	5-Fold Cross Validation with Top-8 Limits . . . . .	61
7.2.4	General Results . . . . .	63
7.3	10-Fold Cross Validation Experiments . . . . .	68
7.3.1	10-Fold Cross Validation with Top-2 Limits . . . . .	68
7.3.2	10-Fold Cross Validation with Top-4 Limits . . . . .	69
7.3.3	10-Fold Cross Validation with Top-8 Limits . . . . .	71



7.3.4	General Results . . . . .	72
7.4	Evaluating Local and Global Modeling Effectiveness . . . . .	76
7.5	Deciding Best k-value on Cross Validation . . . . .	77
7.6	Precision and Recall Values at Top-8 . . . . .	78
8	CONCLUSION . . . . .	81
	REFERENCES . . . . .	85
APPENDICES		
A	WEB LOG'S CAPTURED CONCEPTS and WEB URLS . . . . .	88
B	ADDITIONAL RESULTS and EXPERIMENTS . . . . .	121
B.1	3-Fold Cross Validation Detailed Results . . . . .	121
B.1.1	Ksim Similarity Measures . . . . .	121
B.1.2	Osim Similarity Measures . . . . .	125
B.2	5-Fold Cross Validation Detailed Results . . . . .	129
B.2.1	Ksim Similarity Measures . . . . .	129
B.2.2	Osim Similarity Measures . . . . .	135
B.3	10 Fold Cross Validation Detailed Results . . . . .	141

## LIST OF TABLES

### TABLES

Table 3.1	Sample Transition Probabilities . . . . .	12
Table 4.1	Sample Sessions Transition Table . . . . .	17
Table 4.2	Page Properties in Sample Sessions . . . . .	21
Table 4.3	Avg. Duration Table for Sample Sessions . . . . .	22
Table 4.4	Transition Popularity for Sample Sessions . . . . .	22
Table 4.5	Popular Page Rank Values for Sample Sessions . . . . .	23
Table 5.1	Concept Similarity Example . . . . .	28
Table 5.2	Concept Similarity Example . . . . .	29
Table 5.3	PPR Values and Page Similarities for Sample Sessions . . . . .	32
Table 6.1	Server Log Record's Each Part . . . . .	35
Table 7.1	Precision Recall Calculation Infrastructure . . . . .	45
Table 7.2	P-Value for Each Fold . . . . .	76
Table A.1	Three Level Concepts . . . . .	88
Table A.2	Whole Data Set's Captured Concepts . . . . .	93

## LIST OF FIGURES

### FIGURES

Figure 1.1	General Architecture of Next Page Prediction System . . . . .	4
Figure 3.1	Sample Directed Graph with Access Frequencies . . . . .	12
Figure 3.2	Page Rank Distribution Example [26] . . . . .	13
Figure 3.3	Sample Directed Graph . . . . .	14
Figure 4.1	Directed Web Graph of Sample Sessions . . . . .	18
Figure 5.1	Flow of Concept Determination . . . . .	26
Figure 5.2	Concept URL Relation Example . . . . .	28
Figure 5.3	Directed Graph for Next Page Prediction . . . . .	30
Figure 5.4	Directed Graph for Next Page Prediction . . . . .	32
Figure 6.1	Example Server Log Line . . . . .	34
Figure 6.2	Page Rank Based System . . . . .	38
Figure 6.3	Flow of Page Rank Calculation . . . . .	39
Figure 6.4	Flow of Semantic Tagging . . . . .	40
Figure 6.5	Next Page Prediction System . . . . .	41
Figure 7.1	3-Fold Validation with Top-2 Limit Under Ksim Similarity Metric . . . . .	48
Figure 7.2	3-Fold Validation with Top-2 Limit Under Osim Similarity Metric . . . . .	48
Figure 7.3	3-Fold Validation with Top-2 Limit Local Model Variation Percentage . . . . .	49
Figure 7.4	3-Fold Validation with Top-4 Limit Under Ksim Similarity Metric . . . . .	49
Figure 7.5	3-Fold Validation with Top-4 Limit Under Osim Similarity Metric . . . . .	50
Figure 7.6	3-Fold Validation with Top-4 Limit Local Model Variation Percentage . . . . .	51

Figure 7.7 3-Fold Validation with Top-8 Limit Under Ksim Similarity Metric . . . . .	51
Figure 7.8 3-Fold Validation with Top-8 Limit Under Osim Similarity Metric . . . . .	52
Figure 7.9 3-Fold Validation with Top-8 Limit Local Model Variation Percentage . . .	52
Figure 7.10 3-Fold Ksim Comparison in Global Model . . . . .	53
Figure 7.11 3-Fold Osim Comparison in Global Model . . . . .	54
Figure 7.12 3-Fold Ksim Comparison in Local Model . . . . .	54
Figure 7.13 3-Fold Osim Comparison in Local Model . . . . .	55
Figure 7.14 3-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Global Model . . . . .	55
Figure 7.15 3-Fold Validation ST and PPR Weight Effects on HPR under Osim with Global Model . . . . .	56
Figure 7.16 3-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Local Model . . . . .	56
Figure 7.17 3-Fold Validation ST and PPR Weight Effects on HPR under Osim with Local Model . . . . .	57
Figure 7.18 5-Fold Validation with Top-2 Limit Under Ksim Similarity Metric . . . . .	58
Figure 7.19 5-Fold Validation with Top-2 Limit Under Osim Similarity Metric . . . . .	59
Figure 7.20 5-Fold Validation with Top-2 Limit Local Model Variation Percentage . . .	59
Figure 7.21 5-Fold Validation with Top-4 Limit Under Ksim Similarity Metric . . . . .	60
Figure 7.22 5-Fold Validation with Top-4 Limit Under Osim Similarity Metric . . . . .	60
Figure 7.23 5-Fold Validation with Top-4 Local Model Variation Percentage . . . . .	61
Figure 7.24 5-Fold Validation with Top-8 Limit Under Ksim Similarity Metric . . . . .	62
Figure 7.25 5-Fold Validation with Top-8 Limit Under Osim Similarity Metric . . . . .	62
Figure 7.26 5-Fold Validation with Top-8 Local Model Variation Percentage . . . . .	63
Figure 7.27 5-Fold Ksim Comparison in Global Model . . . . .	64
Figure 7.28 5-Fold Osim Comparison in Global Model . . . . .	64
Figure 7.29 5-Fold Ksim Comparison in Local Model . . . . .	65
Figure 7.30 5-Fold Osim Comparison in Local Model . . . . .	65

Figure 7.31 5 Fold Validation ST and PPR Weight Effects on HPR under Ksim with Global Model . . . . .	66
Figure 7.32 5 Fold Validation ST and PPR Weight Effects on HPR under Osim with Global Model . . . . .	66
Figure 7.33 5 Fold Validation ST and PPR Weight Effects on HPR under Ksim with Local Model . . . . .	67
Figure 7.34 5 Fold Validation ST and PPR Weight Effects on HPR under Osim with Local Model . . . . .	67
Figure 7.35 10-Fold Validation with Top-2 Limit Under Ksim Similarity Metric . . . .	68
Figure 7.36 10-Fold Validation with Top-2 Limit Under Osim Similarity Metric . . . .	69
Figure 7.37 10-Fold Validation with Top-2 Limit Local Model Variation Percentage . .	70
Figure 7.38 10-Fold Validation with Top-4 Limit Under Ksim Similarity Metric . . . .	70
Figure 7.39 10-Fold Validation with Top-4 Limit Under Osim Similarity Metric . . . .	71
Figure 7.40 10-Fold Validation with Top-4 Local Model Variation Percentage . . . . .	72
Figure 7.41 10-Fold Validation with Top-8 Limit Under Ksim Similarity Metric . . . .	72
Figure 7.42 10-Fold Validation with Top-8 Limit Under Osim Similarity Metric . . . .	73
Figure 7.43 10-Fold Validation with Top-8 Local Model Variation Percentage . . . . .	73
Figure 7.44 10-Fold Ksim Comparison in Global Model . . . . .	74
Figure 7.45 10-Fold Osim Comparison in Global Model . . . . .	74
Figure 7.46 10-Fold Ksim Comparison in Local Model . . . . .	75
Figure 7.47 10-Fold Osim Comparison in Local Model . . . . .	75
Figure 7.48 Ksim Standard Deviation in Global Model . . . . .	77
Figure 7.49 Osim Standard Deviation in Global Model . . . . .	78
Figure 7.50 3 Fold Precision Recall Values . . . . .	79
Figure 7.51 5 Fold Precision Recall Values . . . . .	79
Figure 7.52 10 Fold Precision Recall Values . . . . .	80
Figure 7.53 $F_1$ Values of Each Method and Fold . . . . .	80
Figure B.1 First Iteration of 3-Fold Validation with Global Model Under Ksim Similarity	121
Figure B.2 First Iteration of 3-Fold Validation with Local Model Under Ksim Similarity	122

Figure B.3 Second Iteration of 3-Fold Validation with Global Model Under Ksim Similarity . . . . .	122
Figure B.4 Second Iteration of 3-Fold Validation with Local Model Under Ksim Similarity . . . . .	123
Figure B.5 Third Iteration of 3-Fold Validation with Global Model Under Ksim Similarity . . . . .	123
Figure B.6 Third Iteration of 3-Fold Validation with Local Model Under Ksim Similarity	124
Figure B.7 First Iteration of 3-Fold Validation with Global Model Under Osim Similarity	125
Figure B.8 First Iteration of 3-Fold Validation with Local Model Under Osim Similarity	126
Figure B.9 Second Iteration of 3-Fold Validation with Global Model Under Osim Similarity . . . . .	126
Figure B.10 Second Iteration of 3-Fold Validation with Local Model Under Osim Similarity . . . . .	127
Figure B.11 Third Iteration of 3-Fold Validation with Global Model Under Osim Similarity . . . . .	127
Figure B.12 Third Iteration of 3-Fold Validation with Local Model Under Osim Similarity	128
Figure B.13 First Iteration of 5-Fold Validation with Global Model Under Ksim Similarity	129
Figure B.14 First Iteration of 5-Fold Validation with Local Model Under Ksim Similarity	130
Figure B.15 Second Iteration of 5-Fold Validation with Global Model Under Ksim Similarity . . . . .	130
Figure B.16 Second Iteration of 5-Fold Validation with Local Model Under Ksim Similarity . . . . .	131
Figure B.17 Third Iteration of 5-Fold Validation with Global Model Under Ksim Similarity . . . . .	131
Figure B.18 Third Iteration of 5-Fold Validation with Local Model Under Ksim Similarity	132
Figure B.19 Fourth Iteration of 5-Fold Validation with Global Model Under Ksim Similarity . . . . .	132
Figure B.20 Fourth Iteration of 5-Fold Validation with Local Model Under Ksim Similarity . . . . .	133
Figure B.21 Fifth Iteration of 5-Fold Validation with Global Model Under Ksim Similarity	133

Figure B.22Fifth Iteration of 5-Fold Validation with Local Model Under Ksim Similarity	134
Figure B.23First Iteration of 5-Fold Validation with Global Model Under Osim Similarity	135
Figure B.24First Iteration of 5-Fold Validation with Local Model Under Osim Similarity	136
Figure B.25Second Iteration of 5-Fold Validation with Global Model Under Osim Similarity . . . . .	136
Figure B.26Second Iteration of 5-Fold Validation with Local Model Under Osim Similarity . . . . .	137
Figure B.27Third Iteration of 5-Fold Validation with Global Model Under Osim Similarity . . . . .	137
Figure B.28Third Iteration of 5-Fold Validation with Local Model Under Osim Similarity	138
Figure B.29Fourth Iteration of 5-Fold Validation with Global Model Under Osim Similarity . . . . .	138
Figure B.30Fourth Iteration of 5-Fold Validation with Local Model Under Osim Similarity . . . . .	139
Figure B.31Fifth Iteration of 5-Fold Validation with Global Model Under Osim Similarity	139
Figure B.32Fifth Iteration of 5-Fold Validation with Local Model Under Osim Similarity	140
Figure B.33Usage Based Page Rank 10-Fold Validation with Global Model Under Ksim Similarity . . . . .	141
Figure B.34Duration Based Page Rank 10-Fold Validation with Global Model Under Ksim Similarity . . . . .	142
Figure B.35Popularity Based Page Rank 10-Fold Validation with Global Model Under Ksim Similarity . . . . .	142
Figure B.36Semantic Tagging 10-Fold Validation with Global Model Under Ksim Similarity . . . . .	143
Figure B.37Hybrid Page Rank 10-Fold Validation with Global Model Under Ksim Similarity . . . . .	143
Figure B.38Usage Based Page Rank 10-Fold Validation with Local Model Under Ksim Similarity . . . . .	144
Figure B.39Duration Based Page Rank 10-Fold Validation with Local Model Under Ksim Similarity . . . . .	144

Figure B.40 Popularity Based Page Rank 10-Fold Validation with Local Model Under Ksim Similarity . . . . .	145
Figure B.41 Semantic Tagging 10-Fold Validation with Local Model Under Ksim Similarity . . . . .	145
Figure B.42 Hybrid Page Rank 10-Fold Validation with Local Model Under Ksim Similarity . . . . .	146
Figure B.43 Usage Based Page Rank 10-Fold Validation with Global Model Under Osim Similarity . . . . .	147
Figure B.44 Duration Based Page Rank 10-Fold Validation with Global Model Under Osim Similarity . . . . .	148
Figure B.45 Popularity Based Page Rank 10-Fold Validation with Global Model Under Osim Similarity . . . . .	148
Figure B.46 Semantic Tagging 10-Fold Validation with Global Model Under Osim Similarity . . . . .	149
Figure B.47 Hybrid Page Rank 10-Fold Validation with Global Model Under Osim Similarity . . . . .	149
Figure B.48 Usage Based Page Rank 10-Fold Validation with Local Model Under Osim Similarity . . . . .	150
Figure B.49 Duration Based Page Rank 10-Fold Validation with Local Model Under Osim Similarity . . . . .	150
Figure B.50 Popularity Based Page Rank 10-Fold Validation with Local Model Under Osim Similarity . . . . .	151
Figure B.51 Semantic Tagging 10-Fold Validation with Local Model Under Osim Similarity . . . . .	151
Figure B.52 Hybrid Page Rank 10-Fold Validation with Local Model Under Osim Similarity . . . . .	152
Figure B.53 10-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Global Model . . . . .	153
Figure B.54 10-Fold Validation ST and PPR Weight Effects on HPR under Osim with Global Model . . . . .	153



Figure B.55 10-Fold Validation ST and PPR Weight Effects on HPR under Ksim with  
Local Model . . . . . 154

Figure B.56 10-Fold Validation ST and PPR Weight Effects on HPR under Osim with  
Local Model . . . . . 154

# **CHAPTER 1**

## **INTRODUCTION**

Internet users on the World Wide Web (WWW) has increased by the rate of 400% by 2011 [1]. In addition to this, number of web pages that are indexed on the Internet is over 50 billion [2]. According to the studies, the size is doubling itself every six to ten months. Web includes a high volume of data that can be described as a bulk of data, which is unfortunately in a raw format. Since gathering information is an indispensable process in our lives, it is necessary to transform this raw data into information.

Web usage mining is one of the most common approaches for extracting information that is hidden in the web. Web usage mining can be defined as the data mining process that is applied on web page visit specific data. In our research we are focused on combination of web usage mining and structural information of web sites and web site's URLs' conceptual meanings, which can be seen as a hybrid web mining approach, basically depending on web server logs.

Next page prediction in a web site is a widespread and promising research area. Especially for recommendation systems, navigations of users in the web site are used for recommending them new pages. These recommendations are usually specialized in predicting the next page of user. This can be applied on various domains. For instance, in shopping web sites, movie or music web sites such information is very useful for recommending new items by analyzing similar behavior of other users in the web site.

In addition to this, user's typical navigations can be investigated for redesigning a web site. With predicting user's next page navigation, a web site can be redesigned for paying attention to usability. Moreover next page prediction can be used for personalization of web, improving search engines results, caching web pages.

The analysis of user's navigation behavior is usually performed on web site's server logs. There are systems that involve downloading a plug-in to collect navigation information on the client side. However, such systems have several drawbacks such as providing limited additional information and being tied to user's preference to install. For those analysis, in general web usage analyzing is preferred and sometimes it is just supported by a client side plug-in.

In the literature, various techniques have been used in order to analyze the web logs [3, 4, 5, 6] for next page prediction. Data mining techniques are heavily used for this purpose. Clustering, sequence mining, associative rule mining and probability models are some of the popular techniques for predicting the next page of user [7, 8, 9].

Markov models are one of the approaches that is used for calculating the probability of a sequence [10]. They have been studied for random processes and it has been shown that they are well suited for predicting next page of user [9, 11, 12]. In Markov model, using longer sequence of navigation for predicting next page leads to more precise results. On the other hand, using longer navigation sequence increases space complexity. This is the main limitation behind Markov models.

Another preferred approach for predicting user's next page navigation is using Page Rank, which is the algorithm behind Google's search engines [13]. Next page will be the page that has the highest rank in these kinds of systems. The main idea behind Page Rank algorithm is that if one page is popular and it points another page, the page that is pointed by a popular page is more popular than the pointing page. Therefore, in-links of a page's popularity determines the popularity of that page. At this point, popularity can be defined in many different ways. Despite the fact that Page Rank is a promising method for labeling pages that can be used for recommending next page, there is an important disadvantage of page ranking algorithm for this domain. The method produces popular pages in a global context, which does not include user's historical navigation behavior. Ignoring this kind of information causes to produce always very similar results for predicting next page. As a remedy, in [6, 14], it is combined with low-level Markov model (which can also seen as a directed graph). In this work, we extend this hybrid approach with the effect of time spent on the web page, structural information of the page (size of the page) and frequency of transitions. In our work, we define popularity of page transitions and popularity of pages in terms of the frequency of transitions

among pages and frequency of page clicks, respectively.

Therefore, the following factors are considered in this work with page rank calculation;

- visit frequency of the page and transition,
- duration on the page and transition (average time spend on the page),
- size of the page,
- in-link and out-link number of the page

These factors are investigated under two separate algorithms. Duration Based Rank (DPR) algorithm focuses on page duration and size, whereas Popularity Based Page Rank (PPR) algorithm focuses on both page duration and size proportion and frequency of pages.

Another preferred approach for extracting information from web site's is web content mining, which analyzes web site's *content*'s. Data mining and text mining techniques are heavily used for web content mining [15]. In web content mining, web site's internal data and web URL's can be used for next page predictions. Ontology based next page predictions are commonly used [16] in this approach.

Supporting web usage mining with content mining is another approach preferred in several studies in the literature [17, 18]. In our approach we analyze only web URL's content and tag each URL with web site's domain related concepts. In next page prediction, we use this semantic tagging for finding pages that are conceptually similar to a given web URL. For supporting semantic tagging approach, this model uses Popularity Based Page Rank (PPR) as a support argument. In other words, in this novel approach semantic tagging similarity are used mainly for next page prediction and when two candidate page's conceptual similarity is the same, then PPR is used for additional information. This approach is called as Semantic Tagging (ST) approach. Moreover, for investigating the effect of PPR and ST together, we model another approach which combines PPR and ST with equal weights for predicting next pages.

Our recommendation system includes these proposed approaches for evaluating the best method under different criteria. Briefly, the architecture is composed of a sequence of components (see Figure 1.1) that work on the offline process of next page prediction.

In the offline process, Page Finder analyses pages and applies some cleaning operations on the data. After that, Session Finder eliminates sessions that are not fitting for our system. Feature Calculator calculates duration values of pages and transitions, frequency values of pages and transitions and size of pages. Lastly Rank Calculator calculates rank values for PPR, DPR, and UPR ranking algorithms for both local and global models. In the online part of the system Recommender recommends top-n pages related to last visited page that is given to system.



Figure 1.1: General Architecture of Next Page Prediction System

In a nutshell, the contributions of this work are listed below.

- Defining *conceptual similarity* in web page's URLs.
- Defining a page's *popularity* in terms of page's and transition's duration time, length of page and frequency values of pages and transitions.
- Duration Based Page Rank (DPR), a novel page rank algorithm which depends on page visit's duration time and page transition's duration time.
- Popularity Based Page Rank (PPR), a novel page rank algorithm which depends on page visit's duration time and page transition's duration time and combination of it with frequency of page visits and page transitions.

- Semantic Tagging (ST), combination of web content mining and web usage mining. It is a method mainly uses the semantic tagging of URLs and their similarities for next page prediction and as a support method it uses PPR.
- Experimental results on next page prediction with several approaches (UPR, DPR, PPR, ST, HPR) that shows that ST and HPR are promising methods.
- The investigation of the effect of local model (a synopsis of total web graph) and global model (whole web graph) on page rank based next page prediction.

The rest of this report is organized as follows. In the next chapter literature research about next page prediction with web usage mining, web content mining and web structural mining and for each branch related and inspired works are explained. This chapter is followed by background information that includes Page Rank and Usage Based Page Rank (UPR) explanations. In Chapter 4, Duration Based Page Rank (DPR) and Popularity Based Page Rank (PPR) methods are described with examples. In Chapter 5, Semantic Tagging (ST) method and next page prediction mechanism are explained. This chapter is followed by general architecture of the developed system and introduction of the application domain whose web server logs are used in evaluation. In Chapter 7, conducted experiments and their results is expressed. Finally in Chapter 8 we conclude our work with discussion of the results and the future work.

## **CHAPTER 2**

### **RELATED WORK**

In this chapter, we present several studies from literature that are similar to our work in different aspects. In this work, we mainly focus on web mining research area for next page predictions.

Web mining can be defined as application of data mining on one or more Web sites for extracting useful information [19]. Web mining can use Web page documents (web page content, web server logs, hyperlinks etc.) and also web services (query information). Web mining can be divided into three main categories. Those are,

- Web Usage Mining
- Web Structure Mining
- Web Content Mining

In our work, we use all of these techniques. In the rest of this section we will give details about each web mining category with their specified methodologies and several significant literature work.

#### **2.1 Web Usage Mining with Markov Models**

Web usage mining can be defined as analyzing user's navigational behaviors for extracting some useful patterns on user's navigational behaviors on the web site. Usually it retrieves data from web server logs and discovers navigational patterns from that source. Web server logs usually records user's web surfing movements one by one. Web server logs include huge

data about users and their navigational behavior. In addition to this, some sniffing plugins can be added into system for retrieving more information, which is not recorded in web server logs. Some additional movements of the user can be collected from sniffing operations on the web server.

There are several techniques applied on web usage mining. Clustering techniques [20, 21], are heavily used for web usage mining. Moreover associative rule mining [9] techniques are also preferred. In [9], Mobasher et al. work on producing associate rules. In their work, they extract rules for predicting user's next page by using Apriori algorithm. They model a data structure for storing navigational behaviors, which is suitable for recommendation. After analysis of navigational behavior, they use a recommendation engine for next page predictions.

Another method used in next page prediction is employing probabilistic reasoning methods. Especially Markov model and variations of them are used for predicting next page of user's navigation by using historical navigation patterns of users. It depends on the idea that in a sequence of visits of a user, each probability of visiting one page and probability of the binary permutations of this sequence determines the whole sequence's probability [10].

The main disadvantage of Markov models is that, if the order of level increases, which also affects the accuracy of the model, it also increases the space complexity of the model. The probabilities are kept in a huge probability matrix and dimensions can be defined as the combination of pages by the number of the order level. For this reason, some studies aim to reduce the size of Markov model with some pruning methods. The work given in [12] uses Markov model with frequency pruning, confidence pruning and error pruning. It is called *selective Markov model*. In frequency pruning, it is stated that low frequency in training set, tend to predict pages with lower accuracy. For this reason, they apply pruning with a given threshold on page frequencies on the training data set. Similarly in confidence pruning, it is stated that while working with Markov models, from one page node, if probabilities of next pages are closer to each other with very small differences, then the one which has higher probability has *lower confidence*. On the contrary, if the probability of one of the next page is significantly different then this page has *higher confidence*. Therefore in their work they define a confidence threshold for eliminating the pages. Lastly they define the combination of frequency and error pruning, and it is called error pruning method.

Probabilistic reasoning methods have been employed in next page prediction studies in the



literature. In [11], authors define a variable length Markov model depending on the complexity of the problem. They use K-Means clustering for separating navigation paths of users and they group users with similar navigations behaviors. By this way, they decrease the size of the problem, which creates an advantage on calculating Markov models.

The details about Markov models will be explained in Chapter 3.

## **2.2 Web Structure Mining and Page Rank Algorithm**

Web structure mining's aim is to collect data from the structure of a web page by analyzing the links that the page is pointed by and pages that it points to. In web structure mining, link structure is important for the structure mining. One of the most popular methods in web structure mining is page rank algorithms and their variations. Most of the time page-ranking algorithms are used by search engines for finding the most important page related to the search content.

The Page Rank algorithm [5] uses the link structure of pages for finding the most important pages with respect to the search result. The algorithm states that if the in-links (pages that pointed to the page) of a page is important, then out-links (pages that pointed by the page) of the page also become important. Therefore the page rank algorithm distributes the rank value of itself through the pages it points to.

There are models that bias Page Rank algorithm with other type of web usage data, structural data or web contents. In [22] the importance of pages are formulated as the in-link number and out-link number of that page. In their work, it is stated that if one page is in the middle of a dense network, then the importance of it would be more than other pages in the same network.

In [6], Usage Based Page Rank algorithm is introduced as the rank distribution of pages depending on the frequency value of transitions and pages. They modeled a localized version of ranking directed graph with Markov models. They model a synopsis of whole web site specialized for every user, and then they calculate the Usage Based Page Rank value with visit frequency of page's. They prefer a synopsis for giving quick recommendation replies to users. However they do not compare global (as a whole web site) and local (as a synopsis of

web site) modeling's accuracy in their work.

In [3], they modified Page Rank algorithm with considering only the time spent by the user on the related page. However in their work, neither the effect of the size value of pages nor frequency values of page and transition visits are considered. As the frequency value, they calculate the proportion of a specified transition vs. all transition's summation value.

More details about Page Rank algorithm will be explained in Chapter 3.

## **2.3 Web Content Mining and Semantic Web Mining**

Web content mining searches and indexes content of web pages and categorizes them into their concepts. Therefore data mining is done on the content of web pages. There are two main points of view is used in web content mining. One of them is information retrieval and the other one in database view. Web agents and some intelligent tools can be used in information retrieval process and in database view all web page is transformed into database and analysis is performed on database.

Berendt et al. [23] represents a content mining model that maps the web sites' content into a specific ontology, which is created by domain experts. By mapping them into ontology, for next page prediction some inference rules can be obtained from that defined ontology. Despite the fact that the work seems very reasonable and successful, the specification of it makes it inapplicable for using it in a common concept.

Oberle et al. [18] defines a new way for mapping concepts into web sites related to their URLs for using again the inference rules for predicting the next pages. Like Berendt's work, it needs a huge preparation for classifying web URLs and maps them into concepts, which is defined in ontology. In addition to this, this method needs domain experts for creating ontology and mapping URLs to ontology classes.

Moreover, there are some works which is a combination of two or more web mining area for supporting the recommendation systems.

Haveliwala et al. [24] represents a model, which is a combination of semantic information related to web pages with page rank algorithm. In the offline process they calculate page rank

values of different concepts, in the online process they support search results with these rank values. Their work is a combination of web structure mining and web content mining. Their aim is to find more appropriate results for search engines.

In [17], they use both web content mining and web usage mining in a hybrid system. They cluster for usage profiles and content groups concurrently. Then they integrate two different groups for supporting the next page prediction system.

Similarly in [16], they use web usage mining with web content mining together for personalization of web navigations. They extend web server logs with content information, which is called *c-logs*. They use *c-logs* for producing associative rules related to content of pages and user navigations related to these pages.

## **CHAPTER 3**

### **BACKGROUND INFORMATION**

In this chapter, background information about the approaches used for next page prediction in this work are presented. Therefore we will start with Markov model predictions. This will be followed by Page Rank algorithm. It is followed by explanation of Usage Based Page Rank, which is specialized Page Rank algorithm. Lastly, semantic tagging on Web pages is briefly explained.

#### **3.1 Markov Model and Directed Graph**

Whole web site or some local subset of it can be modeled as a directed graph with nodes as web pages and edges as transitions between web pages.

A Markov (chain) model is a mathematical system that undergoes transitions from one state to another, among a finite number of states [10]. It is a random process characterized as memoryless, where the next state depends only on the current state and not on the sequence of events that preceded it. However in the  $k$ th-ordered Markov model transition probabilities can be calculated with previous actions depending on the ordered level of the model.

Hence web page navigations can be modeled as a directed graph which models Markov (chain) model by adding probability values to edge labels of these transactions. In our calculations we used first order Markov model due to the high number of pages that appear in the server logs.

Since Markov model is a directed graph, pages can be assumed as nodes and edges can be assumed as transitions between these pages. Moreover transition probability of states can be

assumed as  $\rho_{i \rightarrow j}$  it the probability of visiting  $i$ th page after visiting  $j$ th page. Hence transition probability of 1st order Markov model of  $\rho_{i \rightarrow j}$  is explained in Equation 3.1, where WS is the whole web site and w is the frequency of web page access.

$$\rho(i, j) = \frac{w_{i \rightarrow j}}{\sum_{k \in WS} w_{i \rightarrow k}} \quad (3.1)$$

In Figure 3.1, a directed graph of 4 pages navigations is modeled. In the edge label, for each transition the frequency of transition is assigned. If we want to calculate the 1st order probability of each page, the results can be found in Table 3.1.

Figure 3.1: Sample Directed Graph with Access Frequencies

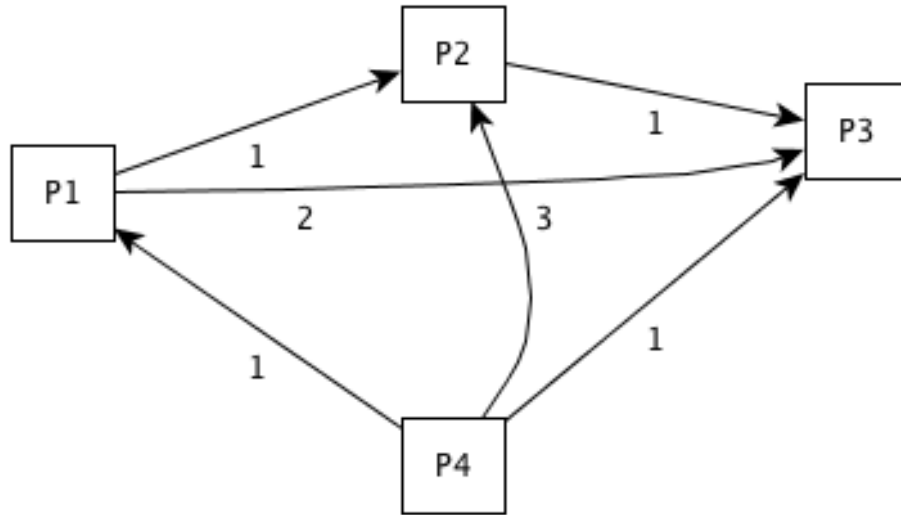


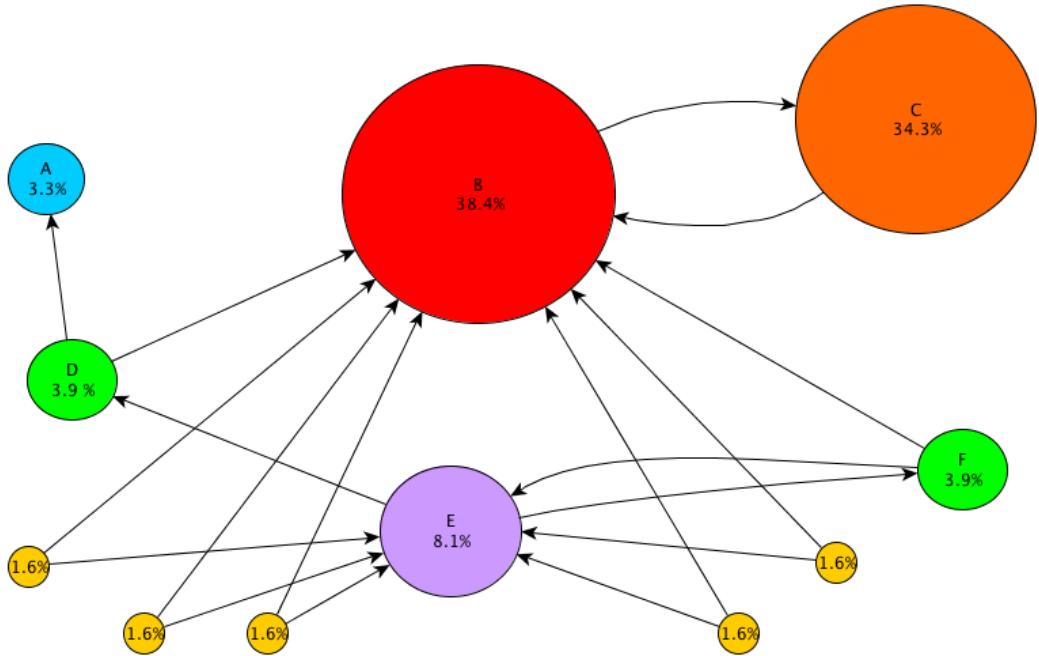
Table 3.1: Sample Transition Probabilities

Transitions	Probability Detailed	Probability
$P1 \rightarrow P3$	$\frac{2}{(1+2)}$	0.66
$P1 \rightarrow P2$	$\frac{1}{(1+2)}$	0.33
$P2 \rightarrow P3$	$\frac{1}{1}$	1.00
$P4 \rightarrow P1$	$\frac{1}{(1+3+1)}$	0.20
$P4 \rightarrow P3$	$\frac{1}{(1+3+1)}$	0.20
$P4 \rightarrow P2$	$\frac{3}{(1+3+1)}$	0.60

### 3.2 Conventional Page Rank Algorithm

Page Rank algorithm [13] models the whole web as a directed graph that keeps nodes as web pages. They use the link structure of pages for determining the importance (rank value) of pages. Google Web search engine [25] mechanism uses Page Rank algorithm for recommending relevant pages to user with ordering them through their rank values. In this algorithm it is stated that if a page has some important in-links to it then its out-links to other pages also become important. In other words if a page is important then pages that it points to are also important. Therefore the algorithm propagates in-links of pages and if the in-links' total is higher then the rank value of it is also higher. In Figure 3.2, the calculation process can be understood clearly. In this example network, rank values are distributed over 100% value. Although page C has fewer in-links then page E, rank value of C is greater than E, which is the explanation of the statement that, whether the in-links of a page is important, then the page is also important.

Figure 3.2: Page Rank Distribution Example [26]



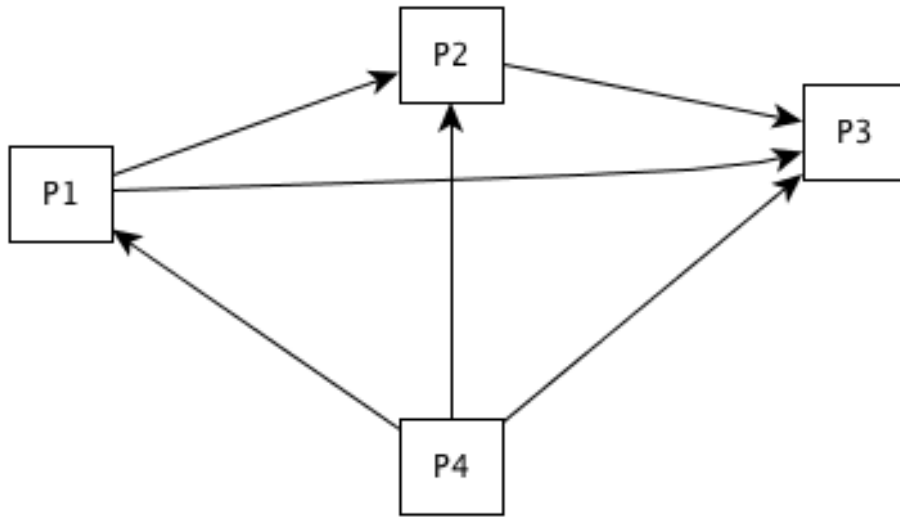
Basic calculation of Page Rank algorithm is given in Equation 3.2.  $IN(v)$  represents the in-links of page  $v$ ,  $OUT(v)$  is the out-links of page  $v$  and  $|OUT_v|$  list the number of out-links of

page  $v$ .

$$PR(u) = \sum_{v \in IN(u)} \frac{PR(v)}{|OUT_v|} \quad (3.2)$$

Page Rank algorithm's basic calculation method can be explained in an example in this section. Suppose in our calculation universe, we have have 4 pages; P1, P2, P3 and P4, respectively. While P1 points to P2 and P3, P2 points to only P3. Lastly P4 points to all pages in our page set. In Figure 3.3, page's navigation relations can be seen.

Figure 3.3: Sample Directed Graph



If we want to calculate the page rank value of P3, first of all we need to calculate page rank value of P1, P2 and P4. Page rank algorithm is implemented for P3 in Equation 3.3.

$$PR(P3) = \frac{PR(P1)}{2} + \frac{PR(P2)}{1} + \frac{PR(P4)}{3} \quad (3.3)$$

In Equation 3.3, it is obvious that in page rank calculation there should be an initial value for all pages in the calculation universe. By using initial values, page rank calculation becomes an iterative process. In iterative calculation method the calculation is implemented with cycles. In the first cycle all rank values may assign to a constant value such as 1, and with each iteration of calculation the rank value become normalized within approximately 50 iterations with  $\epsilon = 0.85$ . Epsilon ( $\epsilon$ ) value will be explained at the rest of this section.

Sometimes the user may not follow a sequenced behavior on Web surfing. They may *jump* to another page that is not linked by her current page. In other words, user may choose another

url without following links and menu bars on her current web site (maybe she writes on the browser a different address or selects a url from her favorites). For this reason, the Page Rank calculation includes a *random surfer jumping* factor on it. With this method, every rank calculation includes not just sequential but also random navigations of user. Random surfer jumping factor also called as dampening factor, symbolizes with  $\epsilon$ . In Equation 3.4, Page Rank algorithm is extended with dampening factor.

$$PR(u) = \frac{(1 - \epsilon)}{WS} + \epsilon * \sum_{v \in IN(u)} \frac{PR(v)}{|OUT_v|} \quad (3.4)$$

Actually in Equation 3.4, the Page Rank algorithm is the interpretation of Markov (chain) model. In the basis of Markov model, random walker principle is applied in order to add a probability of not following the sequential navigation of user. In Markov model, the whole network can be assumed as a huge state space and every state transition is a page navigation. Details about how Page Rank algorithm is interpreted as Markov (chain) model can be found in [27].

### 3.3 Usage Based Page Rank

In [14] Usage Based Page Rank (UPR) is introduced. UPR is a variation of the Page Rank algorithm, based on the visit frequency data obtained from previous users' sessions. Equation 3.5 is given for UPR calculation for  $n$  iterations.  $IN(p_i)$  formulates the set of in-links of page  $p_i$  and  $OUT(p_j)$  formulates the set of out-links of page  $p_j$  and  $w_i$  is the frequency of page  $p_i$  and similarly  $w_{i \rightarrow j}$  is the frequency of page  $p_j$  visit after  $p_i$ .

$$UPR^n(p_i) = \epsilon * \sum_{p_j \in IN(p_i)} \left( UPR^{n-1}(p_j) * \frac{w_{j \rightarrow i}}{\sum_{p_k \in OUT(p_j)} w_{j \rightarrow k}} \right) + (1 - \epsilon) * \frac{w_i}{\sum_{p_j \in WS} w_j} \quad (3.5)$$

In UPR calculation, web pages' frequency is introduced to conventional page rank algorithm. In their work, they use 1st order Markov models to construct the whole web graph with probabilistic reasonings. In this directed graph, they calculate page rank in a *localized* format in order to decrease the respond time of the next page prediction system, which is an online process. This local version of directed graph can be seen as a synopsis of web graph, with a specified depth.



UPR uses random walk behavior of page rank algorithm. It is used for both navigational behavior and random behavior. They use the frequency probability of pages and transitions as well.

### **3.4 Semantic Annotation in Web**

There is a huge amount of data in web pages however very small proportion of it can be processed by machines. Web classifying is one of the methods in order to use this huge data more. Web classifying can be done from content of web pages and also from web URLs. By using semantic terms that is extracted from content or URL of pages, semantic terms can be mapped to web pages, which can be called as semantic annotation. In [28], web classifying is introduced as an extension of text classifying which uses Html pages' content and also web URLs.

In web mining, semantic annotation techniques are heavily used in order to support next page predictions. In general the annotation process can be divided into two main phases. The first step is to establish mappings between existing semantic terms and those need to be annotated in data. In this step, semantic terms and relations between them (rules, hierarchies etc.) are also determined and the second step includes constructing the model that includes the semantic terms and mappings of it.

Semantic mining takes advantage of the semi-structured Web page content. In addition to textual mining, HTML tags and XML markups carry information that concerns layout, navigations and content of pages, which can infer to logical information about web pages easily. Information retrieval techniques and database view [29] of web pages can easily applied on web pages.

## CHAPTER 4

### DURATION AND POPULARITY BASED PAGE RANK

#### 4.1 Defining Sessions

While speaking of user navigations, user sessions should be considered as the basis. In user sessions we analyze user's navigation behaviors (or transitions) in a web site. All user navigations in a web site can be modeled as a directed graph. In Table 4.1 sample user sessions are shown. In transitions column of the table, P symbols are the pages that a user visits in a session with given order.

Table 4.1: Sample Sessions Transition Table

Session ID	Transitions
S1	P1→P2→P3→P4
S2	P2→P4
S3	P1→P2→P4
S4	P2→P3→P1→P2

In Figure 4.1 directed graph of sessions S1, S2, S3 and S4 are modeled. In this graph, node weights and edge weights are page frequencies and transition frequencies, respectively. In order to complete the graph, we add start (S) and finish (F) nodes to the graph, which is an abstraction and those are not actually map to a real page in sessions. We assume that every session starts with start node and finishes with finish node, respectively.

Navigational behaviors on the web page can be modeled as a weighted directed graph that includes pages as nodes and edges as transitions between pages. In addition to this system, frequency of transitions and frequency of pages can be defined in navigational graph by node weights and edge weights. In the rest of this section, two proposed algorithms, Duration Based

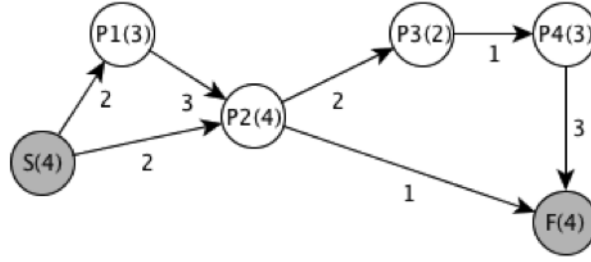


Figure 4.1: Directed Web Graph of Sample Sessions

Page Rank(DPR) and Popularity Based Page Rank(PPR) are presented. In both algorithms, this directed graph is the basis for calculations.

## 4.2 Duration Based Page Rank (DPR)

Distributing rank values of a page to the pages it points to equally is not the best solution for page rank calculation towards next page prediction. In Duration Based Page Rank (DPR) calculation, the distribution simply depends on the duration values of pages and transitions and their web page file size.

Page duration can be defined as the time spent on the page by user after another visit in a given session. Since we want to analyze general behavior of transitions and page rank values, in DPR calculation we use average values of durations. On the other hand, transition duration can be defined as the time spent on two given pages' transitions consecutively. For instance  $P1 \rightarrow P2$  duration can be calculated by searching all  $P1, P2$  transitions in the sessions and retrieving time that is spent on visiting  $P2$  after visiting  $P1$ . Furthermore, we consider the ratio of duration to page size, since in some cases, user spends much time on a web page not for his own interest, but just due to the page size. With the proportion of the two values we aim to focus the real interest of the users on the web pages by considering the file size of them.

General page rank calculation approach adds a random surfer jumping factor to rank value of the page, which means that user may jump to another page that is not a linked navigation [5]. For example, user may write on the internet browser the URL that she wants to go. Hence we can define normal user navigation (sequential) behavior as an edge visit on the graph and

jumping behavior as a node visit on the graph.

DPR calculation uses page duration for random surfing behavior and transition duration for regular visiting behavior of users. The calculation for DPR is given in Equation 4.1.

$$DPR_i = \epsilon * \sum_{x_j \in IN(x_i)} \left[ \frac{DPR_j}{AvgDurationP_{j \rightarrow i}} \right] + (1 - \epsilon) * AvgDuration_i \quad (4.1)$$

### 4.3 Popularity Based Page Rank (PPR)

Popularity Based Page Rank (PPR) calculation, given in Equation 4.2, is modeled in terms of transition popularity and page popularity of the pages that point to (in-links of) the page that is under consideration. In the equation,  $IN(x_j)$  is the set that keeps the in-links of that page.

$$PPR_i = \epsilon * \sum_{x_j \in IN(x_i)} \left[ PPR_j * TransitionP_{j \rightarrow i} \right] + (1 - \epsilon) * PageP_i \quad (4.2)$$

In this equation above, rank distribution of pages in our model depends on the popularity of pages ( $PageP$ ) and transitions ( $TransitionP$ ) that point to that page.

In our model, we define popularity in two dimensions. The first one is page dimension and second one is transition dimension. For both dimensions we define popularity in terms of time user spends on page, size of page and visit frequency of page. Our calculation model is constructed by using coefficients in a different form for assigning rank values to pages than traditional page rank distribution that assigns equal rank values to all in-links of a page.

In popularity calculation, page and transition popularity can be calculated separately but in a similar way. Page popularity is needed for calculating random surfer jumping behavior of the user and transition popularity is needed for calculating the normal navigating behavior of the user. However the main idea is common for finding popularity for nodes and edges. The calculations for transition and page popularity is given in Equation 4.3 and Equation 4.4, respectively.

$$TransitionP_{j \rightarrow i} = Frequency_{j \rightarrow i} * AvgDuration_{j \rightarrow i} \quad (4.3)$$

$$PageP_j = Frequency_j * AvgDuration_j \quad (4.4)$$

The main difference between transition popularity and page popularity can be seen as the focus of their calculation. We start with explaining page dimension and continue with transition dimension.

In Equation 4.5, frequency of page calculation can be found.  $w_j$  is the frequency of visiting  $p_j$  page.

$$Frequency_i = \frac{w_i}{\sum_{p_j \in WS} w_j} \quad (4.5)$$

Average duration calculation that also uses the size of value of pages can be found in Equation 4.6. In this equation,  $d_i$  is the time spend on that page visit until next navigation and  $s_i$  is the size of the page.

$$AverageDuration_i = \frac{\frac{d_i}{s_i}}{\max\left(\frac{d_m}{s_m}\right)}, \text{ where } p_m \in WS \quad (4.6)$$

Finally the open form of page popularity formula can be found in Equation 4.7.

$$PageP_i = \frac{w_i}{\sum_{p_j \in WS} w_j} * \frac{\frac{d_i}{s_i}}{\max\left(\frac{d_m}{s_m}\right)}, \text{ where } p_m \in WS \quad (4.7)$$

Equation 4.8 gives the formula for transition frequency calculation. In this equation,  $w_{j \rightarrow i}$  can be described as the frequency of the transaction. Hence it can be seen as the number of the visits that  $p_i$  after page  $p_j$ . In addition  $OUT(p_j)$  is the pages that point to  $p_j$ .

$$FrequencyP_{j \rightarrow i} = \frac{w_{j \rightarrow i}}{\sum_{p_k \in OUT(p_j)} w_{j \rightarrow k}} \quad (4.8)$$

In Equation 4.9  $d_{j \rightarrow i}$  is duration of the transaction, and  $s_i$  is the size of the transition's result page.  $WS$  is the web page set that includes all pages in the web site. Duration size proportion is inspired from [4] which uses this proportion in a different concept.

$$AvgDurationP_{j \rightarrow i} = \frac{\frac{d_{j \rightarrow i}}{s_i}}{\max\left(\frac{d_{m \rightarrow n}}{s_n}\right)}, \text{ where } p_m \text{ and } p_n \in WS \quad (4.9)$$

In Equation 4.10 transition popularity is defined in terms of transition frequency and duration.

$$TransitionP_{j \rightarrow i} = \frac{w_{j \rightarrow i}}{\sum_{p_k \in OUT(p_j)} w_{j \rightarrow k}} * \frac{\frac{d_{j \rightarrow i}}{s_i}}{\max\left(\frac{d_{m \rightarrow n}}{s_n}\right)}, \text{ where } p_m \text{ and } p_n \in WS \quad (4.10)$$

#### 4.4 PPR and DPR Calculations In Detail

In this section, we present how the given equations are used in the proposed algorithms on a sample case. Since PPR include both frequency, time and page size factors, we present only PPR calculations. In Table 4.2, page id, page size, average duration and frequency value of pages for the sample case are listed.

Table 4.2: Page Properties in Sample Sessions

Page Id	Page Size (byte)	Avg. Duration(ms)	Frequency
P1	1216	297000	3
P2	8103	231000	2
P3	303537	97000	2
P4	9039	10500	3

In Table 4.3 transitions and average transition durations for the sample case are given. Since defining Start(S) and Finish(F) nodes is an abstraction for completing the directed graph, transaction times related to these navigations are not calculated from server logs. In our proposed model, we assigned these transitions the average value of transaction durations.

Assumed valued from the sample case are higher than real values however it should be pointed out that in real data set, these values number are radically less than the values calculated for the sample sessions<sup>1</sup>.

According to these values, popularity rank values of pages can be calculated easily. We show one calculation in detail and give the table of other pages rank values in Table 4.4. Let us calculate *P2* popularity value step by step. From the page popularity equation, popularity of *P2* is calculated as 0.023.

---

<sup>1</sup> In our duration calculations we make 2 iterations. In the first one we calculate exact values of durations and averages of them. In the second iteration we update NA values with average value of durations.

Table 4.3: Avg. Duration Table for Sample Sessions

Transition	Calculated Avg. Duration (ms)	Final Avg. Duration(ms)
S→P1	NA	77000
S→P2	NA	77000
P1→P2	123500	123500
P2→P3	97000	97000
P3→P4	10500	10500
P2→F	NA	77000
P4→P3	NA	77000

$$\begin{aligned}
PageP_2 &= \frac{w_2}{\sum_{p_j \in WS} w_j} * \frac{\frac{d_2}{s_2}}{\max\left(\frac{d_m}{s_m}\right)} \\
&= \frac{2}{(3 + 2 + 2 + 3)} * \frac{28.51}{244.24} \\
&= 0.023
\end{aligned}$$

The values calculated for all transitions in the sample case are given in Table 4.4.

Table 4.4: Transition Popularity for Sample Sessions

Transitions	Frequency	$d/s$	TransitionP
S→P1	2	63.32237	0.50000
S→P2	2	9.50265	0.07503
P1→P2	3	15.24127	0.24069
P2→P3	2	0.31957	0.00336
P3→P4	1	1.16163	0.01834
P2→F	1	1.16667	0.00614
P4→P3	3	1.16667	0.01842

At the end of the first iteration under  $\epsilon = 0.85^2$ , rank values for our sample session is given in Table 4.5. Although we just show the results for one iteration for demonstration purpose, while making next page recommendations, the stability of rank values will be important. This is provided by normalization through further iterations.

---

<sup>2</sup> Commonly in rank calculation experiments  $\epsilon$  value is set to 0.85 and iteration number is set to 50. In our experiments we applied these constants.

In Table 4.5, Popular Page Rank values of pages<sup>3</sup> that are calculated for single iteration are listed.

Table 4.5: Popular Page Rank Values for Sample Sessions

<b>Page</b>	$d/s$	<b>PageP</b>	<b>PPR</b>
P1	244.24342	0.30000	0.45500
P2	28.50796	0.02334	0.141182
P3	0.3195742	0.00026	0.00080
P1	1.16163	0.00143	0.141182
S	NA	0.00323	0.0320

---

<sup>3</sup> As a base of the calculation, we assumed that average values of file size and duration are acceptable for start (S) page of the sessions. So we calculated popularity of S in this table.



## **CHAPTER 5**

### **SEMANTIC TAGGING AND NEXT PAGE PREDICTION**

Web pages include various types of data that can be transformed into useful information. From that perspective, web content mining research area has several sources related to web pages. Web page's text, audio and video objects on the page and even the Uniform Resource Locator (URL) can be used in order to transform data into information which mark the address of a resource on the World Wide Web. In our work, we analyze web URLs in a semantic way in order to obtain useful information for next page prediction of users. Hence, we classify URLs in order to obtain a relationship between web pages in a semantic approach, which is called Semantic Tagging (ST).

In [28], the technique web classification which can be seen as an extension of text classification is explained. Since characterization of web pages are different from normal text documents, the techniques that used to extract information from web pages can be differ from text classification. In addition to structure of web pages, since there can be a meaningful relation between web URLs and web content itself, just text classification may not be enough for covering all properties and characteristics of web pages. In [30], they compare semantic classification of web URLs and conventional content mining approaches in order to observe the effect of web classification by using just web URLs. In the results, it is observed that web classification from only URL is not as effective as content classification, however they observe that web classification can be supported by other mining techniques. From this point of view, we model a novel web classification method and support it with Popularity Based Page Rank (PPR) in order to fill the gap that web classification creates with comparing to web content mining.

In a nutshell, with Semantic Tagging, we analyze each web URL with previously determined

semantic terms and map each term to web URLs with a defined hierarchy.

## 5.1 Semantic Tagging of URLs

Composing semantic information with Web pages is a common approach for supporting web usage mining [18, 23, 24]. In our work, we tag every URL with a specific concept in a concept hierarchy. It should be pointed out that, in our work, we analyze only page URLs for tagging information. In a Web page's URL, usually there exists an information related to semantic meaning of this Web page. In our work, our aim is to find semantic information related to Web address without considering the content of web page. Since in this work's scope we do not focus on Web content mining, we analyze only the URLs of Web pages for supporting next page prediction mechanism with semantic information embedded in Web URLs.

We use 3-level hierarchy for tagging Web pages. In this work, our aim is to explore the effect of concept similarity in next page predictions. Therefore we define a special concept similarity equation. In this equation, we assign each concept level a different weight for measuring the similarity value of each page's conceptual information. In this weight assignment, the main idea is to assign more detailed level higher weight value in order to increase the cumulative concept similarity value. For each level of hierarchy we assign weights with logarithmic distribution starting with 2. In other words, for the first level of detail we assign 2 for  $\lambda_1$ , for the second level of detail we assign 4 for  $\lambda_2$  and for the third level of detail we assign 8 for  $\lambda_3$ .

$$ConceptSim(P_1, P_2) = \sum_{1 \leq n \leq 3} Sim(CS_1, CS_2, n) \quad (5.1)$$

In Equation 5.1, concept similarity is defined by measuring three levels of detailed information related to Web URL where  $P_1$  and  $P_2$  are the pages to be compared for concept similarity.  $CS_1$  and  $CS_2$  are the concept sets related to  $P_1$  and  $P_2$ , respectively.

$$Sim(CS_1, CS_2, n) = \begin{cases} \lambda_n & \exists CS_1[x] \text{ and } CS_2[y] \mid CS_1[x] = CS_2[y], \text{ where } 1 \leq x, y \leq n \\ 0 & \forall CS_1[x] \text{ and } CS_1[x] \mid CS_1[x] \neq CS_2[y], \text{ where } 1 \leq x, y \leq n \end{cases}$$

In semantic tagging process, the first step is to capture the concepts embedded on each Web page URL. After capturing the URLs, detail level of each concept should be determined. After that process, we save each level of concept for calculating the similarity of URLs considering concepts later in an online process of recommendation.



Figure 5.1: Flow of Concept Determination

In Figure 5.1 basic flow of concept determination can be seen. Although it is a manual process, it has a systematic working. In a nutshell, concept determination is started with the prerequisite that semantic terms should already be defined. Then in this process, for each URL, we capture concepts related to this URL. After capturing, we decide the level of each concept in the URL. In that point, the more higher level has more semantic similarity in the web URLs. Finally, we save each level and its mapping with concepts in concept database. In order to expedite this process, we develop a program which saves related URLs with 3-level hierarchy concepts. All captured concepts in three level of details is listed in Appendix-A Table A.1.

As the first step of capturing semantic terms from URLs, we investigate the structure of Web URLs. In each URL, we extract some rules related to each valid value of URLs. The constraints and assumptions considered during semantic tagging process are as follows. Although it is a manual process, it has a systematic working.

- In our methodology, we consider a 3-level of concept hierarchy. However, in some cases, it is hard to capture 3-level concepts. But as a rule of thumb, at least one concept is captured related to URLs.
- Concepts are captured from left to right on the URL text, starting from Level 3 to Level

1. Level 1 has the least and Level 3 has the most detailed information about the Web pages.

- In the calculation of the similarity, since in some cases the captured semantic terms are less than 3, the similarity is searched from the 1st Level to 3rd Level orderly. In comparison, we accept the highest value of coefficients in comparison of different level of concepts.
- The first tag of the URL defines the domain value of the URL and it is captured as the 3rd Level in the hierarchy.
- The second tag of the URL refers to the specific title of the Web page. This concept is captured as the 2nd Level in the hierarchy.
- The third tag of the URL, which is a keyword related to the page, is assigned as 1st Level in the concept hierarchy.

### 5.1.1 Captured Concepts in Experimental Setup

In our experiments we use METU's Web server logs. With the URLs extracted from these server logs, we capture the concepts related to these URLs. Since we prune Web pages with frequency threshold 10, we eliminate non frequent pages from our data set. After pruning, data set contains 628 pages. The full list of semantic tagging on web pages is given in Appendix-A Table A.2.

In Figure 5.2, a sample URL that is taken from our data set, which contains real web URLs. (Full list of web URLs and mapped concepts are given in Appendix-A.) By using this figure, we will investigate how we map */News/thread.php?group=metu.ceng.course.336* URL to concepts in in three-level concept hierarchy. When we look at the URL, it is seen that *ceng.course.336* refers to a course page and course code is *ceng 336*. In addition to this, from the first part of the URL we can notice that the URL belongs to *news* domain. Therefore, as the first step, for each Web page, concepts related to its URL text are captured. After capturing the concepts we decide on the detail level of each concept. In this example, it is obvious that concept *course* is less circumstantial than *ceng 336*. For deciding on the third level of detail, we claim that if the pages are in the same domain, visit chance is more than other related pages in different domains. So, for *news* concept, since it is a domain information about

URL, we decide it to be the third level of concept.

	1st Level	2nd Level	3rd Level
<u>/News/thread.php?group=metu.ceng.course.336</u>	Course	ceng 336	News



  
 cent 336 is the  
course code

Figure 5.2: Concept URL Relation Example

### 5.1.2 Calculating Concept Similarity by an Example

By using three example URLs, we calculate the similarity value with given equations in previous section. Suppose that we want to calculate conceptual similarity of two URLs, from the below list.

- */people/faculty/karagoz/index* ( $P_1$ )
- */~karagoz/ceng302/FurtherDep.ppt* ( $P_2$ )
- */~nihan/ceng302/btrees.ppt* ( $P_3$ )

In Table 5.1, each Web page and its related concepts for each level can be seen.

Table 5.1: Concept Similarity Example

Pages	1st Level Concept	2nd Level Concept	3rd Level Concept
$P_1$	Lecturer	Karagoz	-
$P_2$	Course	Ceng 302	Karagoz
$P_3$	Course	Ceng 302	Nihan

Similarity calculation for  $P_1$  and  $P_2$  is as follows.

1. Level,  $Sim(CS_1, CS_2, 1) = 0$
2. Level,  $Sim(CS_1, CS_2, 2) = 0$
3. Level,  $Sim(CS_1, CS_2, 1) = \lambda_3 = 8$

Similarly for  $P_2$  and  $P_3$ 's concept similarity is as follows.

1. Level,  $Sim(CS_2, CS_3, 1) = \lambda_1 = 2$
2. Level,  $Sim(CS_2, CS_3, 2) = \lambda_2 = 4$
3. Level,  $Sim(CS_2, CS_3, 1) = 0$

And finally, concept similarity of  $P_1$  and  $P_3$  is as follows.

1. Level,  $Sim(CS_1, CS_3, 1) = 0$
2. Level,  $Sim(CS_1, CS_3, 2) = 0$
3. Level,  $Sim(CS_1, CS_3, 1) = 0$

Consequently, final concept similarity values of each web pages can be found in Table 5.2.

Table 5.2: Concept Similarity Example

<b>Pages</b>	<b>Concept Similarity</b>
$P_1$ and $P_2$	8
$P_2$ and $P_3$	6
$P_1$ and $P_3$	0

## 5.2 Next Page Prediction

For predicting the next page, a recommendation set is constructed under the proposed algorithms. The main idea behind predicting next page is to produce recommendations from directed graph that is designed from sessions in web server logs. In the directed graph, for a given depth, recommendation pages are listed and sorted in descending order by calculated rank values. Hence the next page prediction method can be seen as a Markov model that is supported by rank values of pages instead of probabilities. This model can be seen as 1st order Markov model that have a page rank value base.

In our recommendation system, we model three different next page prediction systems for comparing the effect of concept relations and page rank values of pages. The first model, page rank approach, makes predictions only considering the page rank values of pages with our novel Duration Based Page Rank (DPR) and Popularity Based Page Rank (PPR) calculations.

The second model, concept approach, makes predictions by using concept similarity between the current page that is already visited and the next page candidates. This model uses PPR values as an auxiliary method. Finally the third model depends on the both semantic tagging and PPR values of pages, named as hybrid approach.

Consider the example navigation graph given in Figure 5.4 that we mentioned in Chapter 4. If a user visits page  $P_1$ , the recommendation set for depth 2 includes  $P_2$  and  $P_3$  pages, and they will be sorted in descending order with respect to rank or concept similarity values. Therefore the recommendation set will be  $R=\{P_2, P_3\}$  sorted by PPR value in the first prediction model. For concept prediction model, assuming the values that we calculated in previous section, recommendation set will again be  $R=\{P_2, P_3\}$  sorted by conceptual relations. In addition to this, in hybrid approach, this ordering does not change again.

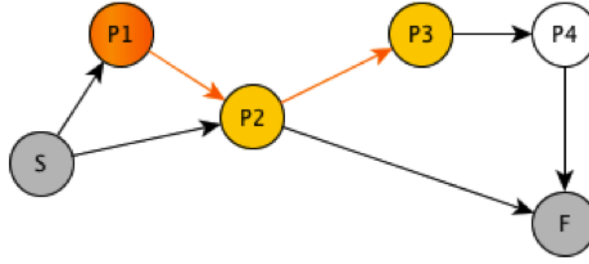


Figure 5.3: Directed Graph for Next Page Prediction

### 5.2.1 Next Page Prediction with DPR and PPR

In order to make rank calculations faster, we record intermediate steps of our calculations to database. Intermediate step values related to rank calculations are, average duration value of pages, average duration values of transitions, page size, frequency value of pages, frequency value of transitions.

After defining sessions and relating them to pages, we calculate average duration values of pages that can be inferred from transition durations that is already recorded. Since for pages that appear at the end of sessions, duration values not be calculated; we assign them average duration values of pages. In addition to this, while analyzing sessions, we calculate transition durations and the size value of pages from web server logs.

Moreover while analyzing sessions, we record page frequencies and calculate transition frequencies. Therefore while analyzing sessions we calculate rank related intermediate values concurrently. Therefore, in our model we recommend set of pages that is sorted by the rank value of the model in descending order on the basis of the pages visited before.

### 5.2.2 Next Page Prediction with Semantic Tagging Approach

While considering conceptual similarity for next page predictions, we analyze each page's URL with methods that are described in the earlier parts of this chapter. After capturing concepts and assigning them to each page, we construct web graph of pages (it is a Markov model for sessions) in both test and training data. From training data, we find the current web page that is visited and we move two steps forward for recording probable next pages for prediction. Following this, we sort these candidates by their conceptual similarities to currently visited page. At this point we use PPR values as an auxiliary mechanism.

For predicting the next page, a recommendation set is constructed under the proposed algorithms. The main idea behind predicting next page is to produce recommendations from directed graph that is designed from sessions in Web server logs. In the directed graph, for a certain depth, pages are listed and sorted in descending order by calculated rank values. Hence the next page prediction method can be seen as Markov model that is supported by page similarity values of pages instead of probabilities. This model can be seen as 1st order Markov model that has a page rank value base.

Consider the example navigation graph given in Figure 5.4. In this example, if a user visits page P1, the recommendation set for depth 2 includes P2, P3 and P4 pages, and they should be sorted in descending order with respect to page similarity values and PPRs.

Assume that, results are already calculated for Popularity Based Page Rank (PPR) and semantic similarity and given in Table 5.3 as an example. With these values, recommendation set is sorted as {P3, P2, P4}. In that point, semantic similarity values are calculated as comparing semantic similarity of the user's current visit with candidate Web pages.

In Semantic Tagging (ST) approach, we use a general conceptual similarity of pages however it occasionally produces the same results especially on semantically irrelevant pages. In this kind of situations, ST method uses Popularity Based Page Rank (PPR) values in order to sup-



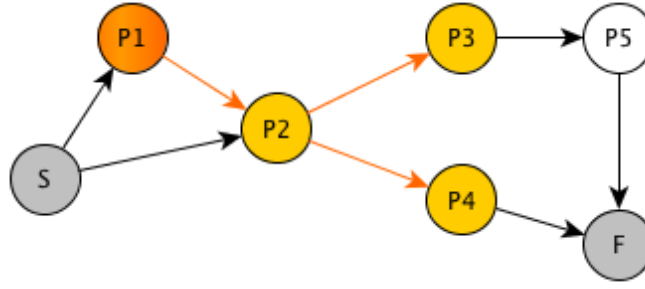


Figure 5.4: Directed Graph for Next Page Prediction

Table 5.3: PPR Values and Page Similarities for Sample Sessions

Page	Popularity Based Page Rank	P1 Semantic Similarity Comparison
P1	0.45500	14
P2	0.141182	4
P3	0.00080	8
P4	0.048265	4
P5	0.00323	0

port conceptual similarity of pages. Therefore Semantic Tagging approach uses conceptual similarity as the basis and it uses Popularity Based Page Rank as a supportive argument.

### 5.2.3 Next Page Prediction with Hybrid Page Rank

In this approach, with comparing Semantic Tagging (ST) approach, we only change the sorting technique of our candidate pages for next page prediction. After capturing concepts of each page's URLs and Popularity Based Page Rank values, we sort our candidate pages for both PPR and conceptual similarities with equal weights.

As an example, assume that we have two lists of the same elements that is sorted by two different orders; semantic similarity values with comparing to last visited page is  $PS_1$  and  $S$  is the all unique page set respectively and  $n$  is the top- $n$  limit of recommendation sets. The algorithm of sorting pages can be seen below.

---

**Algorithm 1** Next Page Prediction with Hybrid Page Rank Approach

---

**for all**  $P$  in  $S$  **do**

**for**  $i = 1 \rightarrow n$  **do**

**if**  $P = PS_1[i]$  **then**

$index_1 \leftarrow i$

**end if**

**if**  $P = PS_2[i]$  **then**

$index_2 \leftarrow i$

**end if**

**end for**

$index_3 \leftarrow (index_1 + index_2)/2$

$map[P] \leftarrow index_3$

**end for**

{After that, sort by index values from map descending}

{If the index values are equal, use PPR values of pages.}

---

In Hybrid Page Rank (HPR) approach, the conceptual similarity and PPR values are used to sort candidate recommended pages with equal weights. However in ST method, conceptual similarity has a priority in the calculation. In HPR, we remove this priority in order to observe the variation of accuracy in recommendations.

## CHAPTER 6

### GENERAL ARCHITECTURE OF THE NEXT PAGE PREDICTION SYSTEM AND APPLICATION DOMAIN

In this chapter, next page prediction system that is developed for this research is presented. In addition, software components of the next page prediction system and their interface relationships are also explained. It is followed by the introduction of the application domain and in this context, web server logs of the domain that are used in our system are introduced and explained.

#### 6.1 Data Set of the Application Domain

In our experimental evaluations, we use METU's web server logs from 29/May/2010 to 18/Feb/2011. We choose METU's web server logs since it includes all data related to our work (i.e. transferred data size, visited URL, client name and user agent name).

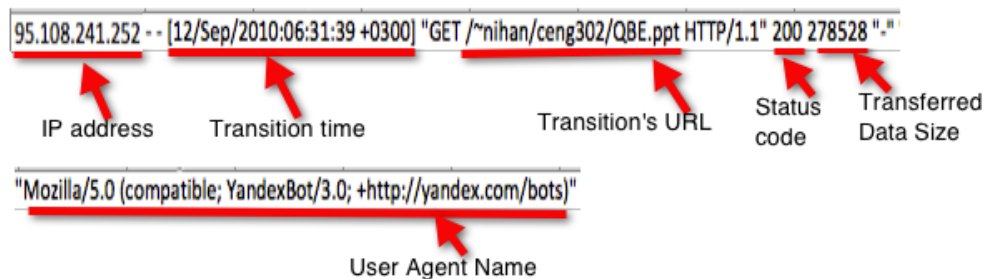


Figure 6.1: Example Server Log Line

Every movement is recorded as a line in the log file and every transition of users' is recorded into web server logs. As shown in In Figure 6.1, in every line of record, the IP address of the request, transition's time, target URL, status code of request, transferred data size and user agent name is kept. In Table 6.1 detailed information about each part of the log record is explained.

Table 6.1: Server Log Record's Each Part

Part Name	Explanation
IP Address	This is the IP address of the machine, that makes the HTTP request. In every request the IP address is recorded by the server into logs. For defining sessions, we use IP address for determining if the transitions are in the same session or not.
Transition's Time	For every hit, the time is recorded in the server log and the time format is <i>dd/MMM/yyyy:hh:mm:ss Z</i> . Every <i>Z</i> value is standardized through RFC 822 time zone. In our work, we use the recorded time for inferring the time duration of each visit by subtracting each consecutive pages in the session.
Transition's URL	Every requested URL with HTTP protocol produces a status code referring the request. Status code has a value with three digits and every number defines the status of request. In a nutshell, all codes can be generalized into four main categories. These categories are listed below.
Status Code	Status code can be a number 2XX, 3XX, 4XX or 5XX. 2XX code refers to success, 3XX refers to redirection, 4XX refers to client error and 5XX refers to server error. In our work, we eliminate requests with 4XX or 5XX codes, since they are interrupted with errors.
Transferred Data Size	Every request transfers an amount of byte for producing web page in the client's web browser. The transferred byte size is recorded in every page request.

Table 6.1: (continued)

User Agent Name	Client's internet browser and operating system information is recorded with related request in the web server log. Usually it includes internet browser's version, name and operating system of the client. With IP address, this information is also included for finding sessions.
-----------------	--

## 6.2 Analyzing Server Logs

### 6.2.1 Automatic Pruning of Web Server Logs

In our work, since we work with 6.5 million of different web pages, we automatize elimination of pages from sessions. We define a set of pruning rules and the useless pages to be eliminated from the database are determined with respect to these rules.

In the pruning step, we eliminate some pages which we analyze that related URLs would not help on next page predictions. For instance home page ("'/index.php'"), log-in, log-out operations and frame downloads related to main pages are such pages. In this scope, we eliminate URLs with ".png, .gif, .jpg" extensions and similarly download related URLs are also pruned automatically. Moreover, we eliminate *news* related frame pages including "*left, right*" keywords. We define a set of banned words for pruning URLs that include them. Those banned words are "*login, logout and download*". Lastly we used frequency pruning [12] and we eliminate web pages having frequency below 10.

### 6.2.2 Session Identification

In our work, we determine some rules in order to identify a session. First of all, in a session, whole transitions must be recorded with the same IP and user agent name (described in the previous section). In our work, we read all web server logs and keep them in map like data structure, which has a key of combination of IP number and user agent name and which stores

URL of each mapping. After reading all records in the web server logs, page elimination is started and irrelevant and useless pages are filtered from page map. The pruned page map is the input of the session identification process.

In our work, sessions are defined as the page transitions that occur with the same IP and user agent between 30 minutes. Moreover, if the idle time in one page is more than 10 minutes, this ends the session and starts another one including the next visited page. Pseudocode of the session identification can be found below.

---

**Algorithm 2** Session Identification

---

```

m ← 0
for all K in eliminatedPageMapKeys do
    sameSession ← true
    i ← 0
    for all P in eliminatedPageMapKeys.get(K) do
        if (time(P) − time(P1) < 30) and (duration(P) < 10) then
            sessionm[i] ← P
            i ← i + 1
        else
            m ← m + 1
            i ← 0
            sessionm[i] ← P
        end if
    end for
end for

```

---

### 6.3 General Architecture of the Next Page Prediction System

In our research, we develop several modules combining with each other for constructing the whole system named *next page prediction system*. The general architecture of the system contains three main sub systems; page rank based system, concept tagging system and recommender system.

### 6.3.1 Page Rank Based System

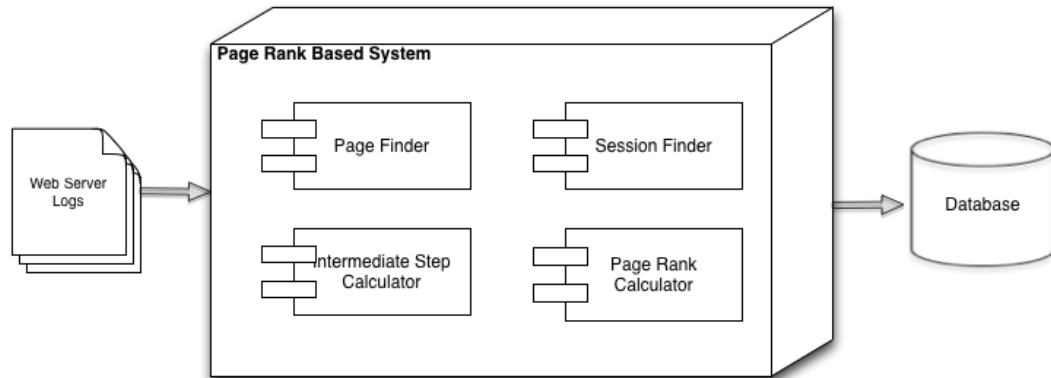


Figure 6.2: Page Rank Based System

In Figure 6.2, general architecture of page rank based system can be seen. In this system, *PageFinder* eliminates useless pages before session identification and records pruned pages into database. After that, *SessionFinder* identifies sessions and sessions are recorded into database. Session identification is followed by page rank calculation operations. In this step, *PageRankCalculator* constructs 1st order Markov model (directed graph) of the whole transitions, which is recorded previously. After constructing the directed graph, Usage Based Page Rank (UPR), Duration Based Page Rank (DPR) and Popularity Based Page Rank (PPR) values are calculated for both local and global model (Details about each page rank calculation algorithm can be found in Chapter 4 and global, local models can be found in Chapter 3).

In Figure 6.3, the general flow of the page rank calculation method can be seen. In sequence, web server logs are read, all the pages that are read from web server logs are recorded into database with their frequency values. Then pages are pruned from *dirty* and infrequent data. After that sessions are identified from web server logs. In this step all transitions related to web navigations are extracted from web server logs and recorded into database. It is followed by recording sessions that are identified from web server logs. Finally page rank values are calculated from sessions and transitions.

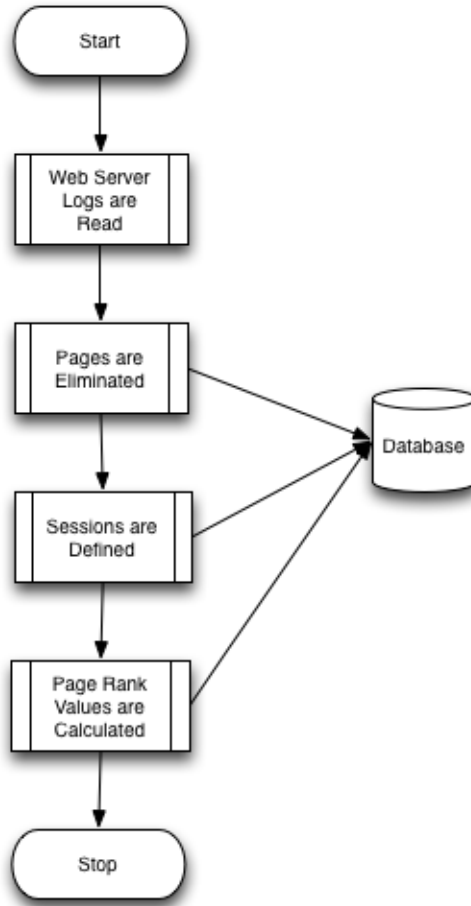


Figure 6.3: Flow of Page Rank Calculation

### 6.3.2 Semantic Tagging Based System

Semantic Tagging based system can be seen as an extension to the previous *Page Rank Based System*. This system uses Page Rank based system for page elimination and session identification. After that, a manual process for *semantic tagging* is started in order to annotate each page to related concept with three levels of detail (More information about concept capturing and mapping can be found in Chapter 5).

In Figure 6.4, the basic flow of semantic tagging approach is shown. As can be seen in this figure, in order to calculate semantic similarity of pages, PPR values of pages should be already defined. For this reason, first of all page rank calculation should be calculated. Hence semantic tagging approach is dependent to page rank calculation in next page prediction.



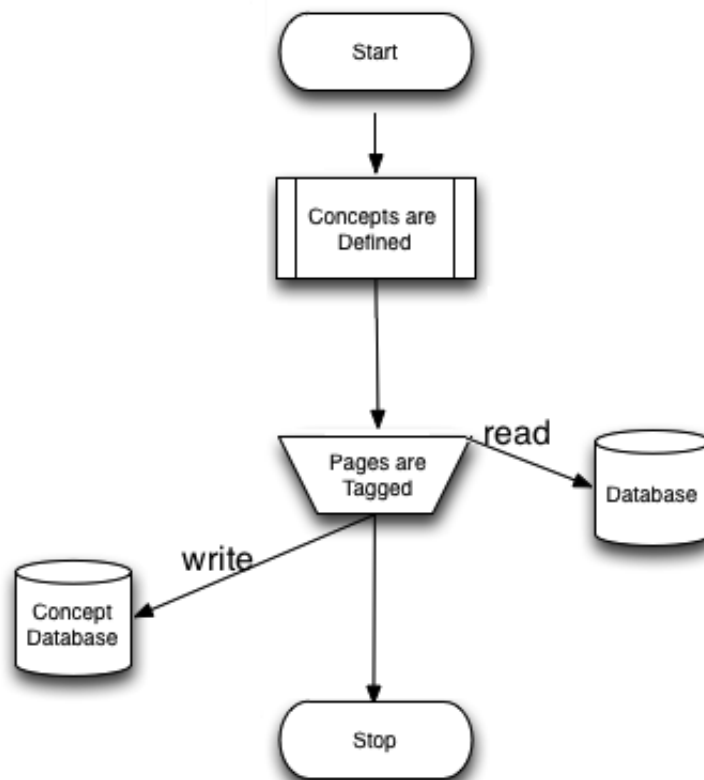


Figure 6.4: Flow of Semantic Tagging

However, in semantic tagging process, it is only needed to read pruned pages from database, which is already recorded during page rank calculation. In this phase of process, it is not necessary for semantic tagging. After reading pages from database, all pruned pages are tagged to concepts in three level of detail. As a prerequisite, semantic terms in the domain should be extracted.

### 6.3.3 Next Page Prediction System

*Next Page Prediction System* uses both *Page Rank Based System* and *Semantic Tagging Based System*. Its aim is to produce for 1st level Markov model visits of users suitable next pages with the previously mentioned algorithms and their variations. First of all, it uses the *Page Rank Based System* for predicting pages only produced from pure page rank algorithms (UPR,

DPR and PPR). Moreover this system uses *Semantic Tagging Based System* for producing results only produced from semantic similarities of each page respectively. And finally, it produces hybrid version of Popularity Based Page Rank (PPR) and conceptual similarity of pages and produces next page candidates with two measures that are weighted equally.

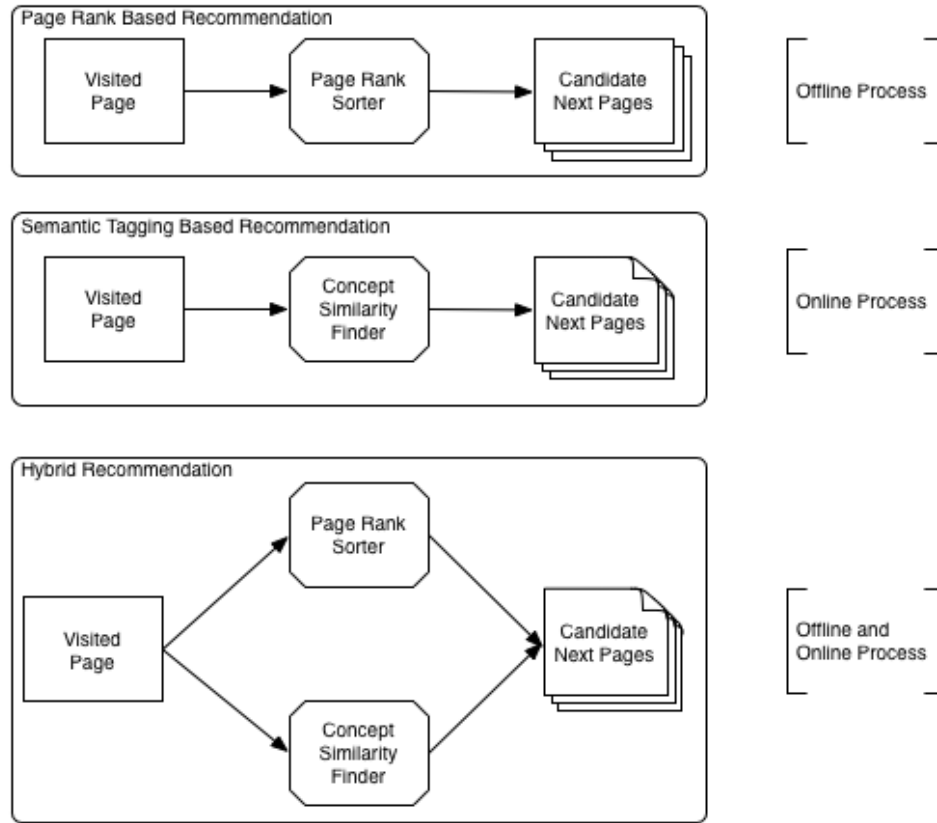


Figure 6.5: Next Page Prediction System

Three general approaches are summarized in Figure 6.5. The first partition of the system needs offline calculation of each (UPR, DPR and PPR) page rank calculation for finally recommending suitable next page candidates. On the other hand, semantic tagging system does not need a previous calculation of conceptual similarities since it is a less complex operation than calculating page rank of each page.

## CHAPTER 7

### EXPERIMENTS AND EVALUATION

In our experimental evaluations, we use METU's web server logs from 29/May/2010 to 18/Feb/2011. In the raw data there are 5.168.361 unique page URLs.

In the experiments, we analyze and compare the accuracy of page ranking models (UPR, DPR, PPR), Semantic Tagging based (ST) and lastly Hybrid Page Rank (HPR) next page prediction models. In our experiments basically we use the evaluation method employed in [6]. They use two different data sets in their evaluations and they employ *holdout* method for validating their estimation model.

In holdout model, data dependency is very high and *unfortunate* splitting of training and test set may cause misleading error rate. In our experiments, we prefer k-fold cross validation method, which supports the independency of test and training data. Moreover k-fold cross validation method allows all data to be in test and training partition. More details about k fold cross validation and hold-out methods can be found in [31].

Basically in k-fold cross validation method, the data set is divided into k parts. After partitioning, one of the k parts is selected for being test data and all other parts are grouped into training data and for each partition experiments are performed and results are recorded. Finally the estimation result ( $E$ ) is calculated as the Equation 7.1.

$$E = \frac{1}{K} * \sum_{i=1}^K E_i \quad (7.1)$$

In k-fold cross validation method, choosing the *best* k value is another problem. In [32], they investigate the k values and evaluate each of them with bias and standard deviation parameters. Moreover they emphasize that data variation and overlapping on test and training data sets

may determine the  $k$  value and sometimes lower  $k$  values can be preferred. In addition to this, in [33], they observe that the best  $k$  value is 5 instead of commonly preferred value 10. In the last part of this chapter, the standard deviations of each fold values with different methods is presented.

In order to perform with the best fold number in our cross validation method, we run our tests for 3-fold, 5-fold and 10-fold. In this data set, we observe that since unique page number is very high, dividing the test data set in a small partition as 1/10, results with the lowest accuracy results in all models. When the partition size gets smaller, test and training data tend to have less number of common elements, which drops the accuracy. For this reason, we prefer to evaluate results from 3-fold and 5-fold cross validation. However, we give results run on 10-fold cross validation in Appendix-B.

Moreover, for each folding experiments, for evaluating the effect of Popularity Based Page Rank (PPR) and Semantic Tagging (ST) prediction methods in the hybrid approach, we run some extra tests and evaluations considering the two of them. In addition to this, for each folding, we evaluate the local and global modeling of the page ranking for each method (UPR, DPR and PPR).

Since our aim is to find the *best* next page predictions for current visit of the user, each next page prediction model produces recommendations ordered by the each model's specified methodology (i.e. popularity based page rank value, conceptual similarity etc.). At this point, recommending only one page is not the common behavior of the next page prediction system. For this reason, we want to investigate the effect of the recommendation limits of the system. Therefore, in each validation method, we perform our experiments for 2, 4 and 8 next page candidates.

For every validation method, the data is pruned and preprocessed under the criteria mentioned in Chapter 6. After preprocessing, for calculating the page rank values, the formulas given in Chapter 4 and Chapter 3 are applied under  $\epsilon=0.85$ , which results 0.15 jumping factor and 50 iterations [6]. Rank values are calculated for all three algorithms (Usage Based, Duration Based and Popularity Based Ranking algorithms). It should be pointed out that, these calculations are performed for both global and local model with depth 2.

After page pruning, sessions are identified from web server logs. In these sessions, we pro-

duce a directed web graph of test data in order to produce real transition values and to compare them with the predictions. In every evaluation we pick one page in the directed graph that have 2 or more nodes that it points to. Then for that page, every algorithm produces recommendation sets.

In comparing the predictions with the real page visits, there are two similarity algorithms that are commonly preferred [6, 14, 24, 3] for finding similarities of two sets. In our experiments, we also use these methods. The first one is called *Osim* [24] algorithm, which calculates the similarity of two sets without considering the ordering of the elements in the set. It focuses on the number of common elements of two sets with a limit value. The limit value can be seen as the top-n recommended pages for a visited page. The equation of *Osim* algorithm is defined in Equation 7.2, where  $A$  and  $B$  are the sets to be compared, that have the same length and  $n$  is the top-n value of comparison. The similarity value range is [0-1] and 1 denotes maximum similarity.

$$Osim(A, B) = \frac{|A \cap B|}{n} \quad (7.2)$$

As the second similarity metric we use *Ksim* similarity algorithm, which concerns Kendall Tau Distance [24, 6] for measuring the similarity of next page prediction set produced by training data set and real page visit set on the test data. Kendall Tau Distance is the number of pairwise incompatibility between two sets. It is also titled as *bubble sort distance* since it is equivalent to the number of swaps for making the two lists in the same order by using bubble sort algorithm. In this similarity metric, as the distance increases, similarity decreases. *Ksim* similarity calculation is given in Equation 7.3. Sometimes the compared sets may have different lengths. The lengths of the sets are equalized, by utilizing the union set, as shown in Equation 7.3.

$$\begin{aligned} \delta_1 &= A \cup B - A \text{ and } \delta_2 = A \cup B - B \\ A' &= A \text{ followed by } \delta_1 \text{ and } B' = B \text{ followed by } \delta_2 \text{ then,} \end{aligned}$$

$$Ksim(A, B) = 1 - \frac{\tau \text{ distance}(\delta'_1, \delta'_2)}{|A \cup B| * (|A \cup B| - 1)} \quad (7.3)$$

$\tau$  distance has come from the Kendall Tau distance algorithm mentioned before.

In addition to *Osim* and *Ksim* measurements, for a certain subset of our experiments we calculate precision and recall values of recommendations with top-8 limit values. Precision is the

probability of randomly selected recommended page is relevant to real visit of user and recall is the probability that a randomly selected visited page is retrieved in the recommendation. In Table 7.1, the general form of precision recall calculation page sets are given in a statistical classification.

Table 7.1: Precision Recall Calculation Infrastructure

	<b>Relevant Pages</b>	<b>Irrelevant Pages</b>
<b>Recommended</b>	A	B
<b>Not Recommended</b>	C	D

With this classification given in Table 7.1, precision and recall Equations 7.4 and 7.5 are given.

$$Precision = \frac{A}{(A \cup B)} \quad (7.4)$$

$$Recall = \frac{A}{(A \cup C)} \quad (7.5)$$

In [34], effectiveness  $E$  formula is given in Equation 7.6 where  $P$  is the precision value,  $R$  is the recall value.

$$E = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} \quad (7.6)$$

$$F_\beta = 1 - E \text{ and } \beta = \frac{1}{1 + \beta^2}$$

The most preferred value [34] of measuring the accuracy of recommendation with  $\beta = 1$ . Hence the formula becomes for  $F$  value as  $F_1$  and it is the harmonic mean of precision and recall values. In our experiments we calculate precision recall values of each fold in average and we also calculate  $F_1$  values of each methodology with each fold.

In a nutshell, in our experiments we test methodologies below.

- Usage Based Page Rank (UPR)
- Duration Based Page Rank (DPR)
- Popularity Based Page Rank (PPR)
- Semantic Tagging (ST)

- Hybrid Page Rank (HP)

next We test our methodology with two modeling.

- Local Model (Synopsis of a Web Graph)
- Global Model (Whole Web Graph)

We validate our data with 3 variations.

- 3-Fold Cross Validation
- 5-Fold Cross Validation
- 10-Fold Cross Validation

In addition to this, for each cross validation, we investigate the *optimum* decision point for Hybrid Page Rank (HPR)’s proportions on Semantic Tagging (ST) and Popularity Based Page Rank (PPR).

Moreover we apply student t-test for analyzing whether the effect of global and local modeling is statistically significant and for deciding the best fitted  $k$  value in cross validation, we calculate each methods standard deviation value for all iterations.

Lastly we give results of precision recall values of each methodology and fold value in scatterplot and we give  $F_1$  values of each methodology in each fold.

In the rest of this chapter, each of these experiments is explained.

## 7.1 3-Fold Cross Validation Experiments

For training set after pruning, we have 4577 page values and for test set after pruning, we have 1650 page values. After identifying sessions, we obtain 320 training sessions and 180 test sessions.

In our experiment setup, we make separate iterations with top-2, top-4 and top-8 recommendation comparisons that are measured by Ksim and Osim with global and local ranking methods. The results of the experiments for the next page prediction accuracy for three different

page ranking algorithm, semantic tagging method and hybrid page rank method under *Ksim* and *Osim* similarity metrics and local and global models are given in the rest of this section.

In each next page prediction experiment, we produce results for both global and local models. In [6], authors prefer local model for improving the calculation time of each page's rank in online mode. Since in our system we need to prepare data before the recommendation system and we improve the calculation time of page rank values with the design of our *Intermediate Step Calculator* (More details can be found in Chapter 5), we can freely choose either local or global model. In [6], the basic drawback of global model is reported as its inefficiency. However, in this work, we decrease the global model calculation time by storing the intermediate results. By this way, we can benefit from the global model without increasing the time cost. For this flexibility on choosing models, we investigate the *effectiveness* of both models for producing results. Again in our experiments, we prepare the system for top-n limits with *Ksim* and *Osim* similarity metrics. For each fold, we group local and global models in the following figures.

### 7.1.1 3-Fold Cross Validation with Top-2 Limits

The experiments are run in both global and local context of the model with top-2 limits under *Ksim* similarity metric. The results are presented in Figure 7.1. The same experiments are also evaluated with *Osim* similarity metric, which can be found in Figure 7.2.

From these results, it is observed that, for both global and local context models, Semantic Tagging (ST) next page prediction system and Hybrid Page Rank (HPR) system make more accurate recommendations than other methods. In comparison, both semantic tagging and hybrid page rank methods improve next page predictions under *Ksim* similarity in top-2 limit in the average of local and global models by 38%. Under *Osim* similarity comparison, Semantic Tagging (ST) method is 28% more effective and Hybrid Page Rank (HPR) method is 52% more effective than Usage Based Page Rank method.

In order to investigate the effect of local model in comparison to global model while calculating next page predictions, we calculate the change percentage of local model to global model. It is the percentage of change in local model with comparing the local and global model difference. In Figure 7.3, recommendations with limit value 2 are evaluated with two similarity



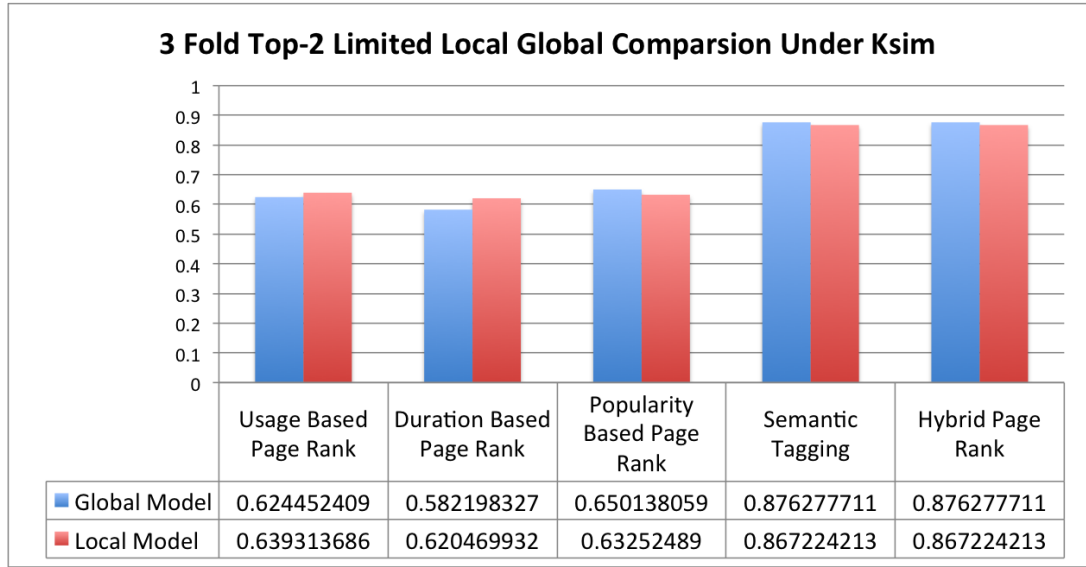


Figure 7.1: 3-Fold Validation with Top-2 Limit Under Ksim Similarity Metric

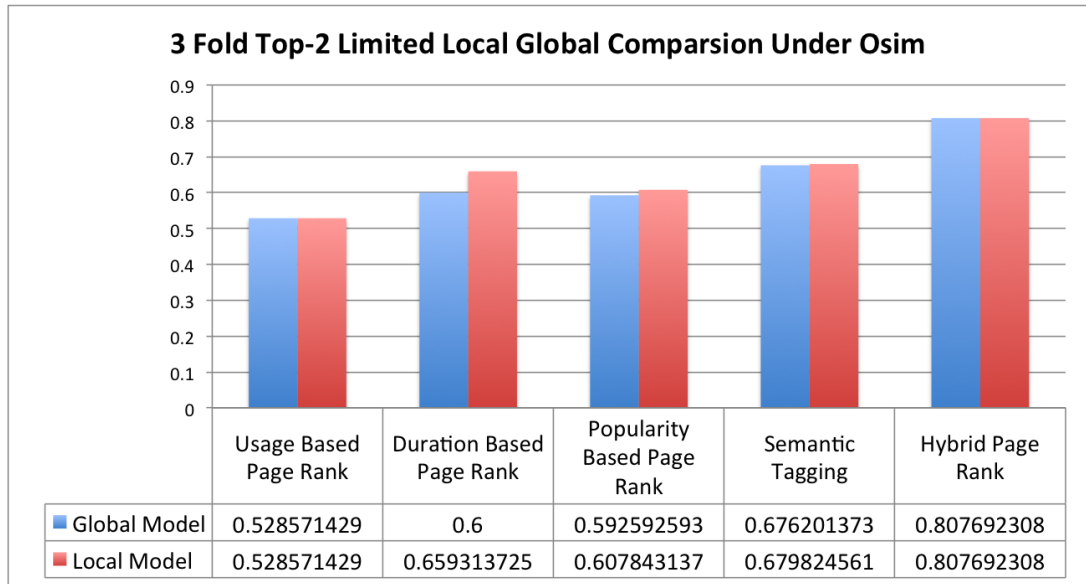


Figure 7.2: 3-Fold Validation with Top-2 Limit Under Osim Similarity Metric

metrics.

### 7.1.2 3-Fold Cross Validation with Top-4 Limits

Top-4 experiments that are conducted in both global and local context of the model under *Ksim* similarity metric can be seen in Figure 7.4. Moreover the same experiments are also

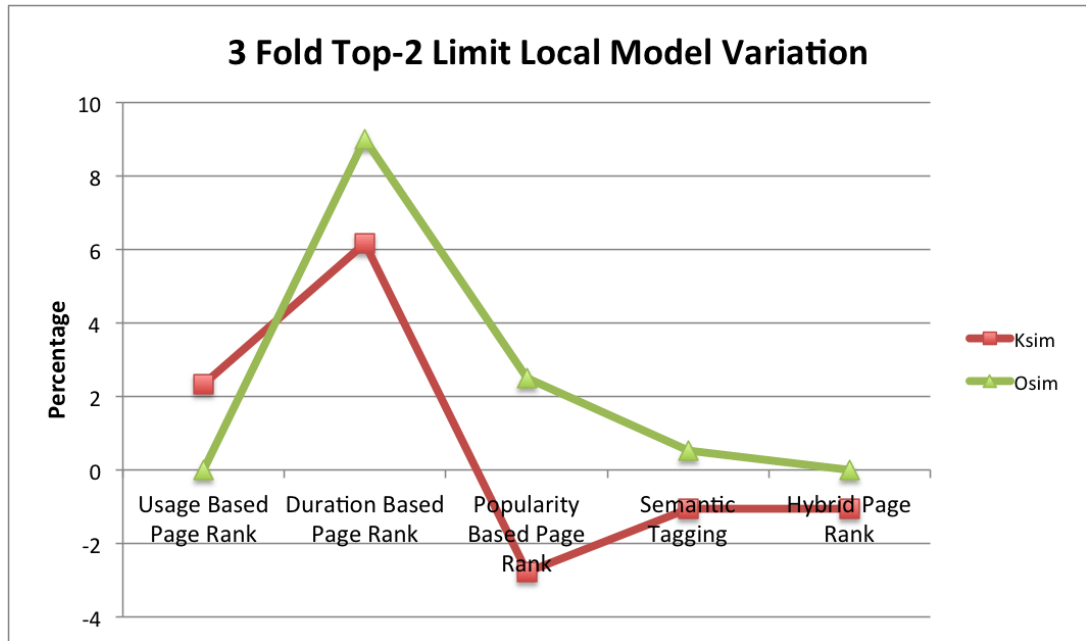


Figure 7.3: 3-Fold Validation with Top-2 Limit Local Model Variation Percentage

evaluated under with *Osim* similarity metric, which can be found in Figure 7.5.

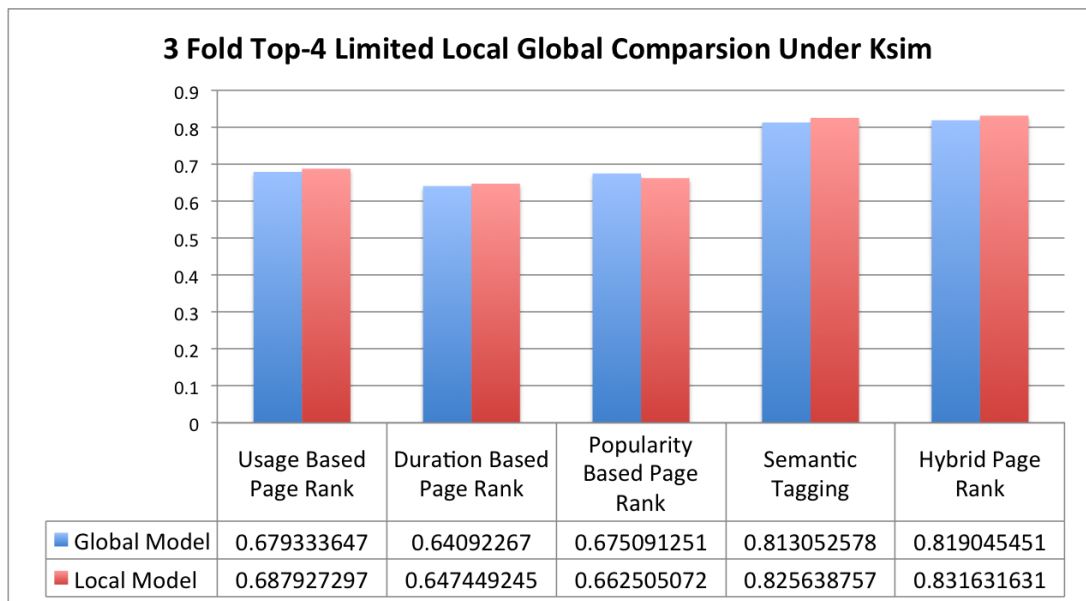


Figure 7.4: 3-Fold Validation with Top-4 Limit Under Ksim Similarity Metric

From these results, it is observed that, for both global and local context models, our Semantic Tagging (ST) next page prediction system and Hybrid Page Rank (HPR) system make more accurate recommendations than other methods. In comparison, both semantic tagging and

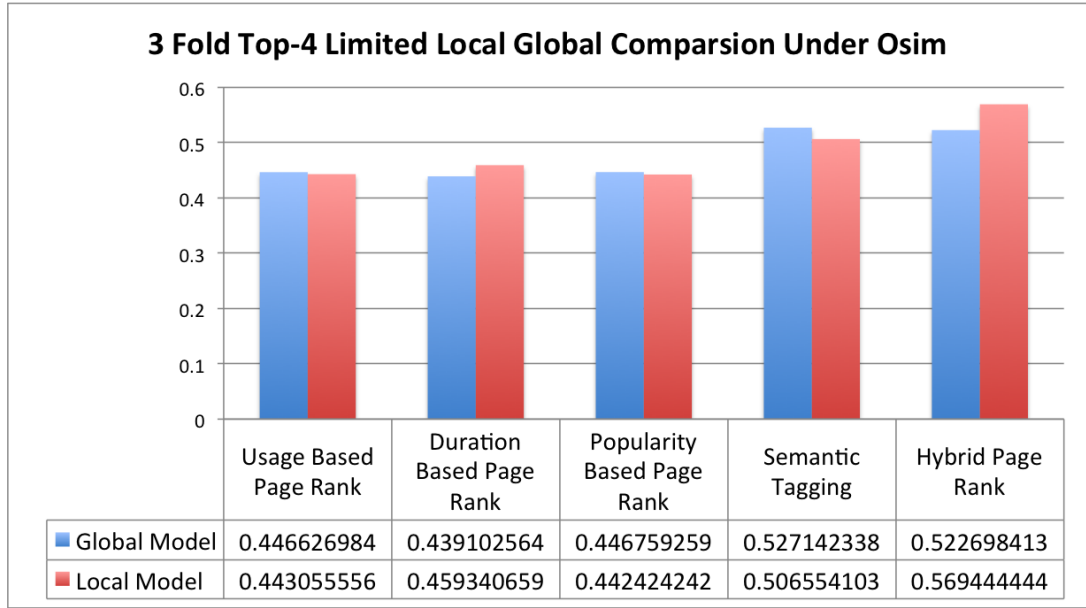


Figure 7.5: 3-Fold Validation with Top-4 Limit Under Osim Similarity Metric

hybrid page rank methods improve next page predictions under *Ksim* similarity metric in top-4 limit by 20%. Under *Osim* similarity comparison, Semantic Tagging method is 16% effective and Hybrid Page rank method is 22% effective than usage based page rank method.

In order to investigate the effect of local model to global model with top-4 next page predictions, we calculate the change percentage of local model to global model. In Figure 7.6, recommendations with limit value 2 are evaluated with two similarity metrics.

### 7.1.3 3-Fold Cross Validation with Top-8 Limits

In top-8 limit of experiments, as being a maximum value, the general aim is to investigate the limit value behaviors of each next page prediction methodology. Like other experiments, they run in both global and local context of the model with top-8 limits under *Ksim* and *Osim* similarity metric. The results can be found in Figure 7.7 and Figure 7.8.

From these results, it is observed that, for both global and local context models, our Semantic Tagging (ST) next page prediction system and Hybrid Page Rank (HPR) system make more accurate recommendations than other methods. In comparison, both semantic tagging and hybrid page rank methods improve next page predictions under *Ksim* similarity metric in

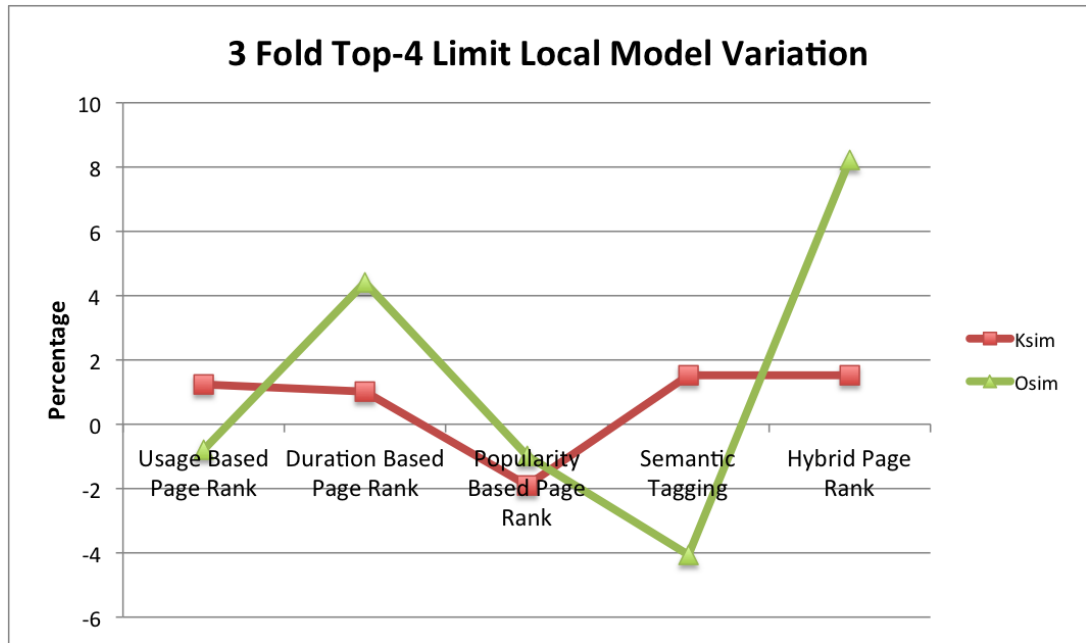


Figure 7.6: 3-Fold Validation with Top-4 Limit Local Model Variation Percentage

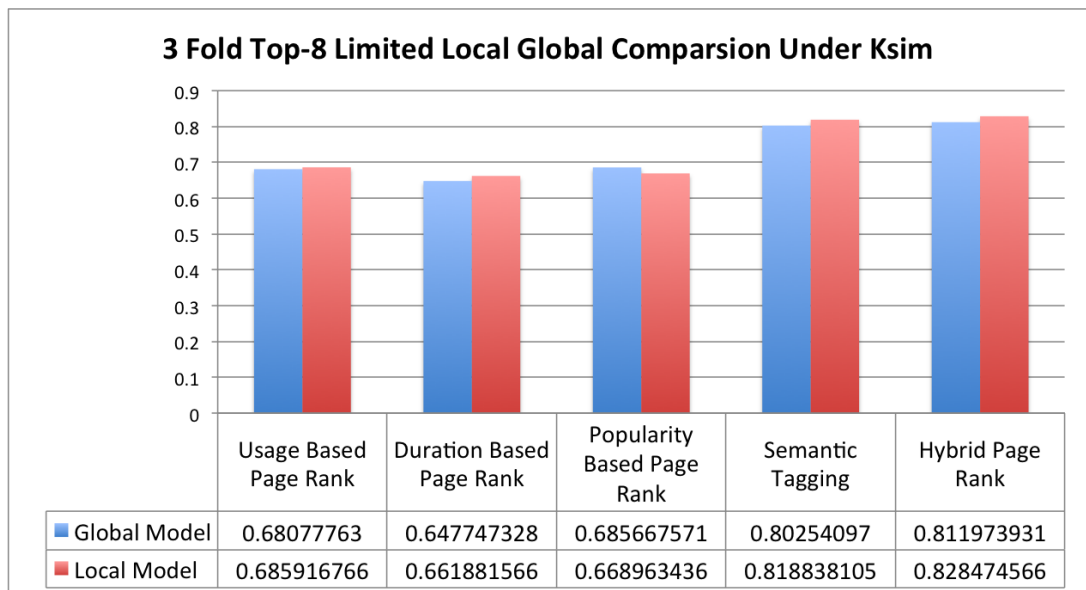


Figure 7.7: 3-Fold Validation with Top-8 Limit Under Ksim Similarity Metric

top-8 limit by 19%. Under *Osim* similarity comparison, Semantic Tagging method is 33% effective and Hybrid Page rank method is 28% effective than usage based page rank method.

In order to investigate the effect of local model in comparison to global model with top-8 next page predictions, we calculate the change percentage of local model to global model. In

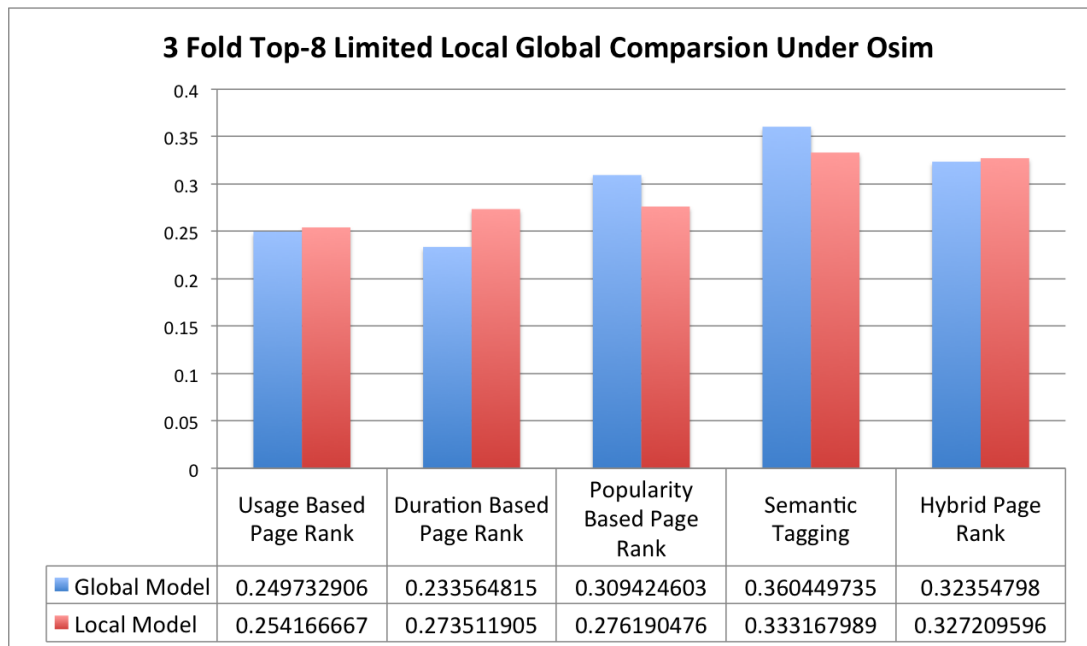


Figure 7.8: 3-Fold Validation with Top-8 Limit Under Osim Similarity Metric

Figure 7.9, recommendations with limit value 2 are evaluated with two similarity metrics.

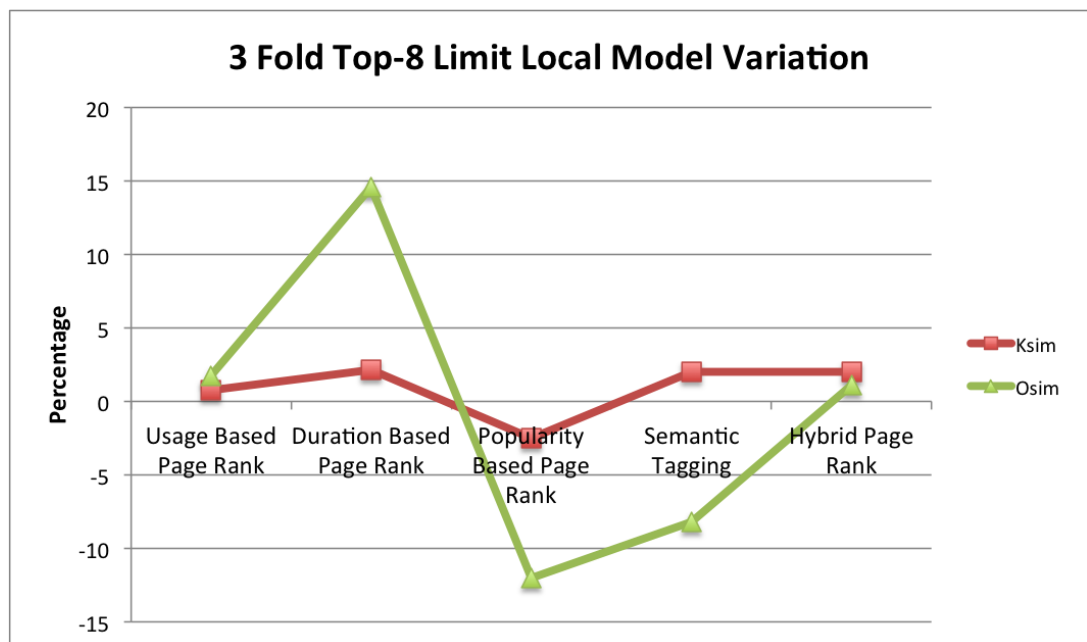


Figure 7.9: 3-Fold Validation with Top-8 Limit Local Model Variation Percentage

#### 7.1.4 General Results

In conclusion, for each top-n limits, Semantic Tagging and Hybrid Page Rank methods are at least 20% effective than previous methods with 3-fold cross validation. Moreover, it is observed that especially with *Osim* similarity metric, which only considers the common elements of the real data set and recommendation set, the effectiveness of all models are getting lower with higher top limit values, which seems very reasonable. In next page prediction systems, the main aim is to find actual next candidate of the user, instead of recommending him a bulk of page set.

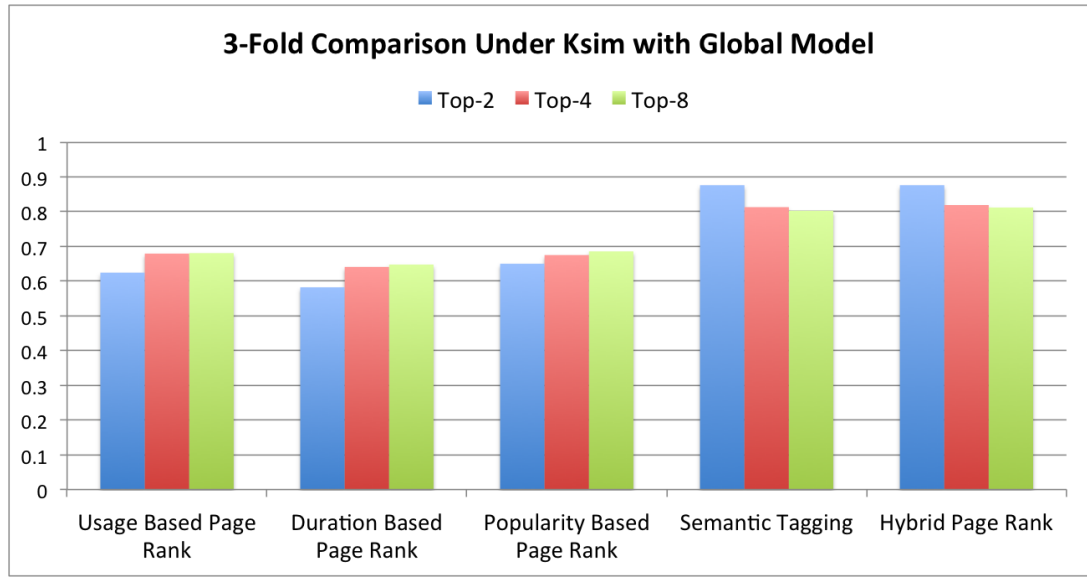


Figure 7.10: 3-Fold Ksim Comparison in Global Model

In our general experiments in Hybrid Page Rank method, we always assign Popularity Based Page Rank and Semantic Tagging values with equal weight (0.5 - 0.5) for constructing the model. In this work, we also investigated the variation of weights in Hybrid Page Rank model and evaluated the accuracy. In the experiments we generate 9 different models and assign values starting with proportion of 0.1 to Semantic Tagging and eventually 0.9 to Popularity Based Page Rank. Then we iterate each run with increasing the effect of Semantic Tagging with 0.1 and eventually increasing Popularity Based Page Rank weight with 0.1. The results for each similarity metric and model can be found in Figure 7.14, Figure 7.15, Figure 7.16 and Figure 7.17.

Under *Ksim* similarity, (0.2 - 0.8) pair can be preferred which is closer to Popularity Based

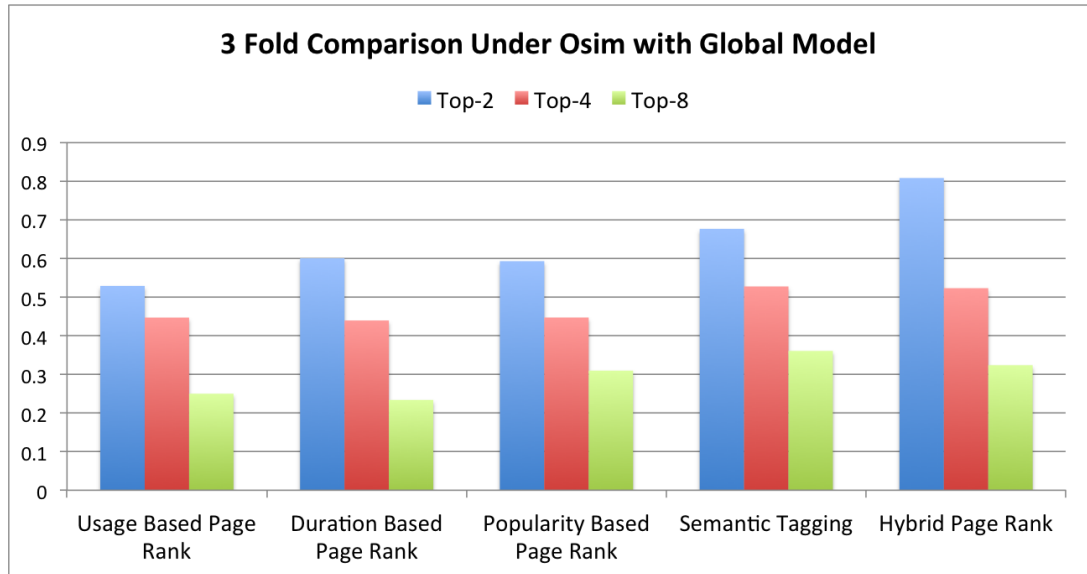


Figure 7.11: 3-Fold Osim Comparison in Global Model

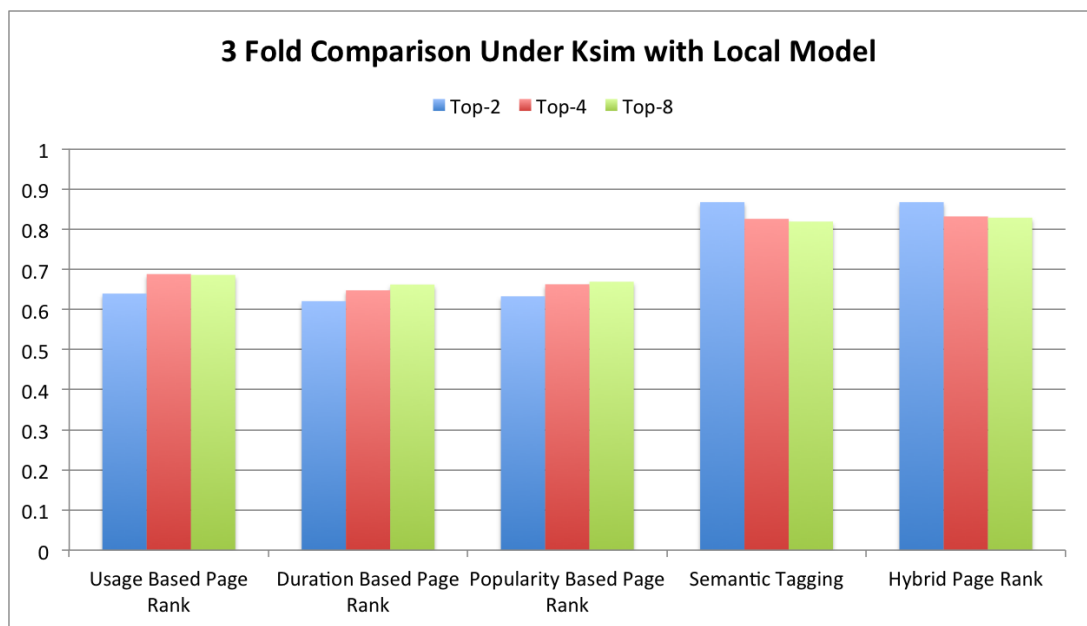


Figure 7.12: 3-Fold Ksim Comparison in Local Model

Page Rank however, in *Osim* similarity (0.9 - 0.1) pair can be preferred for both global and local models and this behavior is common for each top-n limit values. As a result, since *Osim* similarity measures only the common values of real data set and recommendation set of user, a hybrid model closer to Semantic Tagging can be chosen.

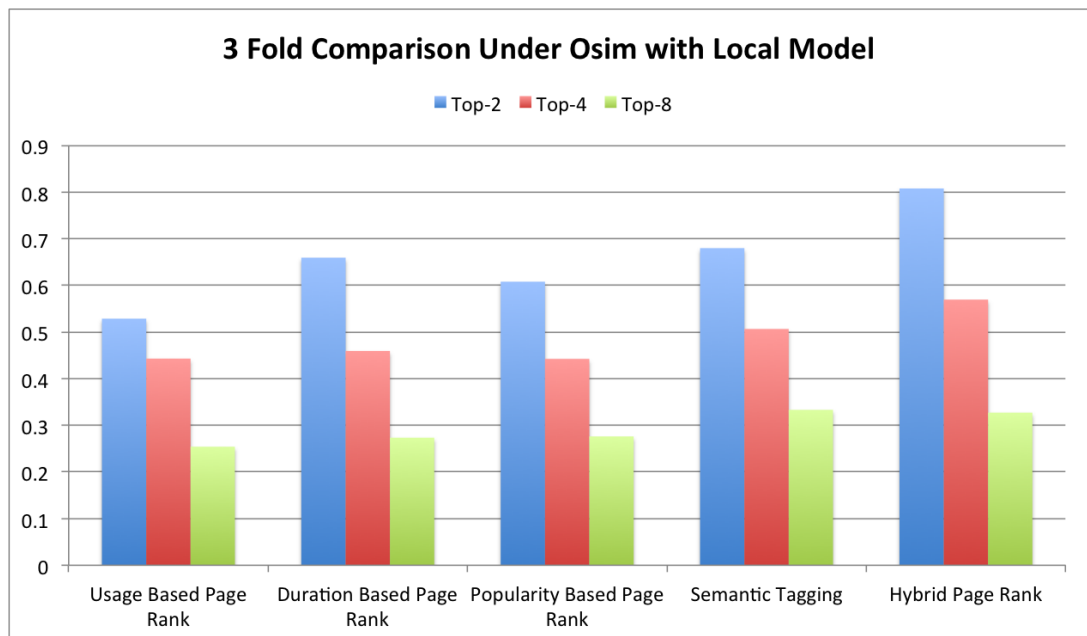


Figure 7.13: 3-Fold Osim Comparison in Local Model

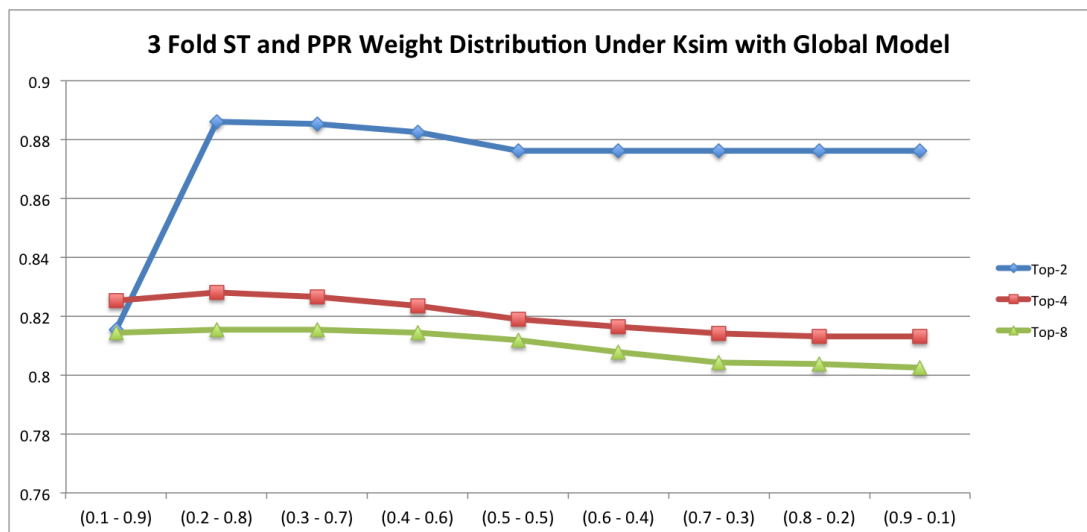


Figure 7.14: 3-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Global Model



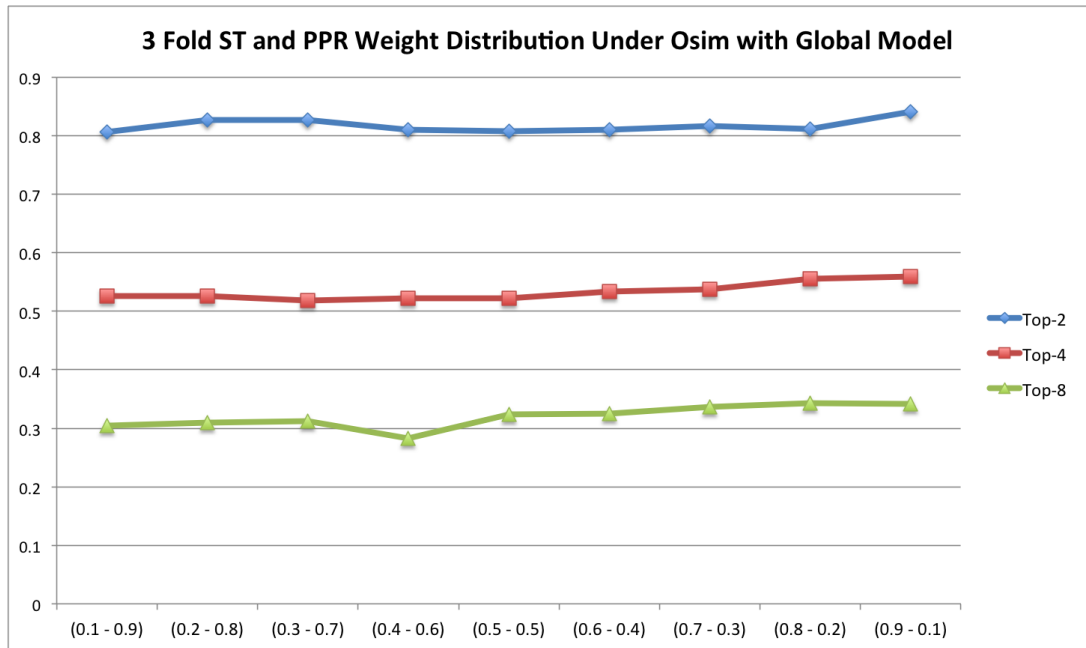


Figure 7.15: 3-Fold Validation ST and PPR Weight Effects on HPR under Osim with Global Model

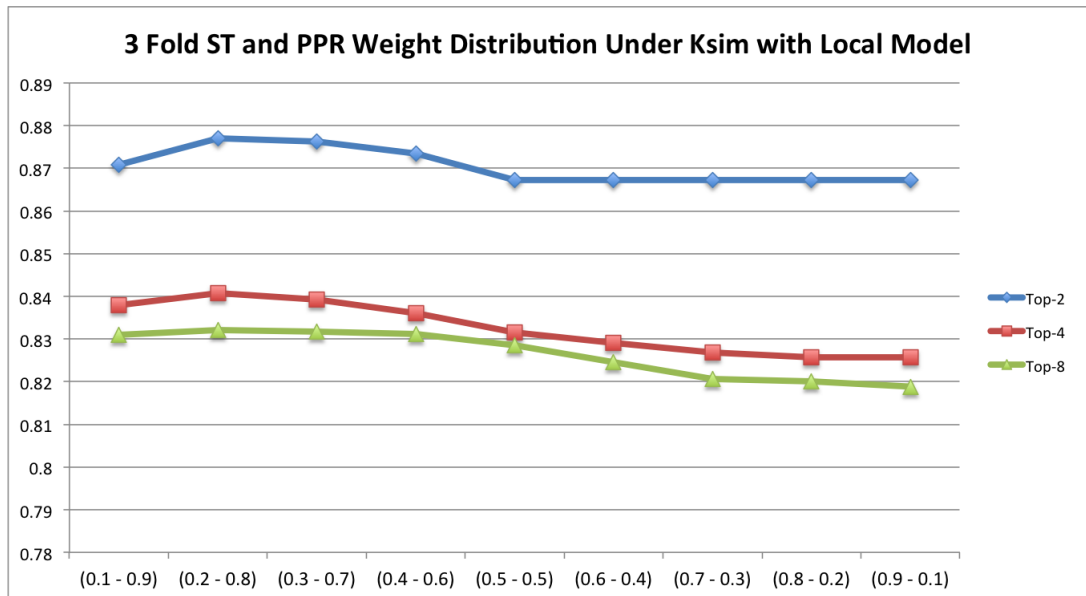


Figure 7.16: 3-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Local Model

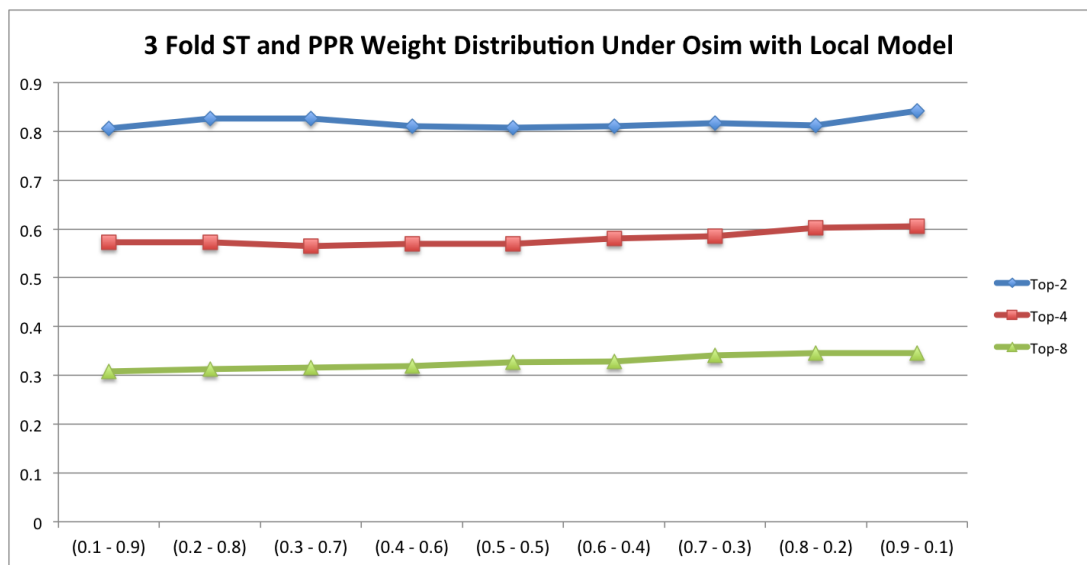


Figure 7.17: 3-Fold Validation ST and PPR Weight Effects on HPR under Osim with Local Model

## 7.2 5-Fold Cross Validation Experiments

Secondly we run our experiments with 5-fold cross validation and for this reason we divide data into five parts and in each iteration we pick one of the not previously chosen for test data. Moreover in our setup, we make separate iterations with top-2, top-4 and top-8 recommendation comparisons that are measured by *Ksim* and *Osim* with global and local ranking methods. All five next page prediction model's (UPR, DPR, PPR, ST and HPR) accuracy under *Ksim* and *Osim* similarity metrics and local and global models are given in the rest of this section.

### 7.2.1 5-Fold Cross Validation with Top-2 Limits

The experiments that are conducted in both global and local context of the model with top-2 limit under *Ksim* similarity metric can be seen in Figure 7.18. Moreover the same experiments are run and evaluate with *Osim* similarity metric, which can be found in Figure 7.19.

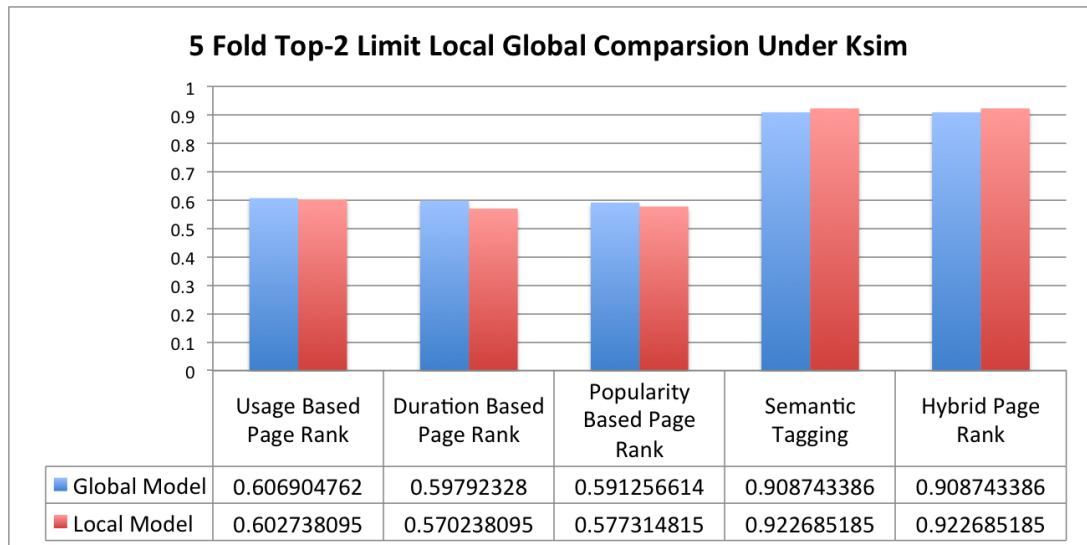


Figure 7.18: 5-Fold Validation with Top-2 Limit Under Ksim Similarity Metric

Under *Ksim* similarity, both Semantic Tagging and Hybrid Page Rank approaches improved the previous works by nearly 50% for both global and local model. However by measuring the *Osim* similarity, the effectiveness of the Semantic Tagging and Hybrid Page Rank become under 3%. In addition, under *Osim* similarity metric, it is observed that all methodology's results are getting closer to each other, so they lose their identifications. Their predictions

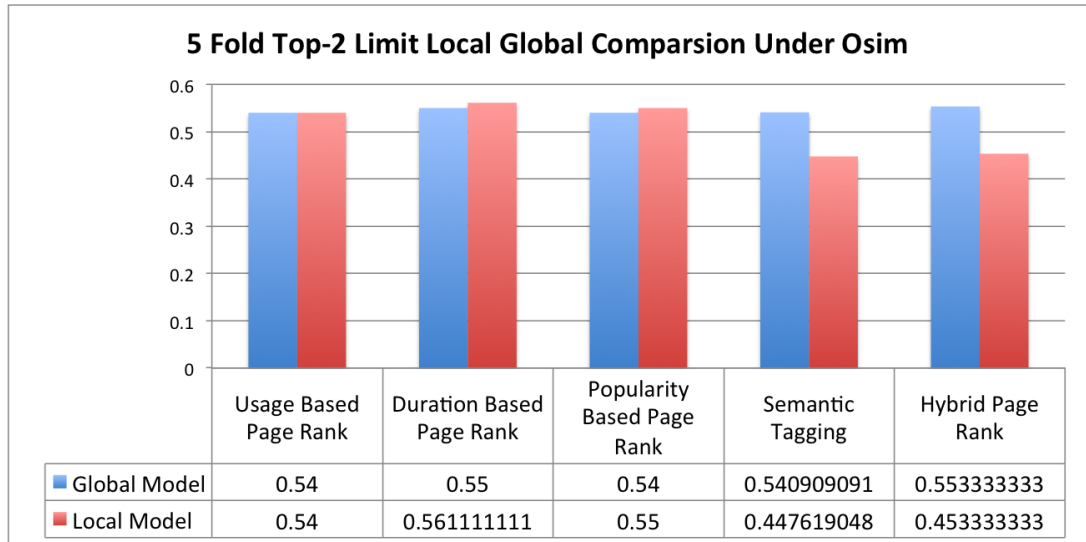


Figure 7.19: 5-Fold Validation with Top-2 Limit Under Osim Similarity Metric

are getting closer due to the lack of common data in test and training data under 5-fold.

In order to investigate the effect of local model to global model while calculating next page predictions, we calculate the change percentage of local model to global model. In Figure 7.20, recommendations with limit value 2 are evaluated with two similarity metrics.

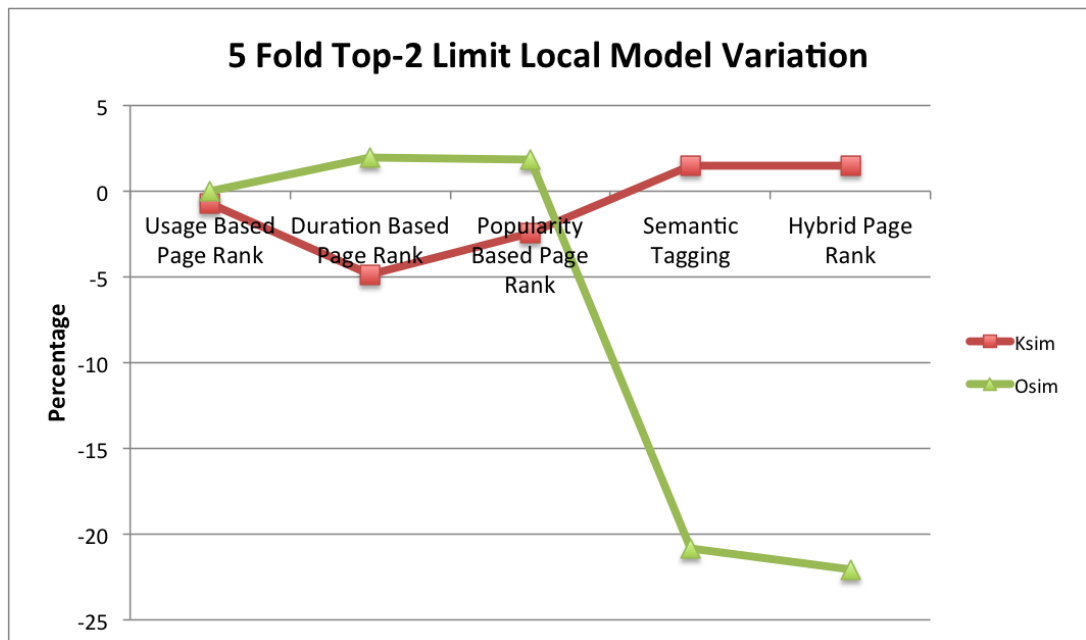


Figure 7.20: 5-Fold Validation with Top-2 Limit Local Model Variation Percentage

## 7.2.2 5-Fold Cross Validation with Top-4 Limits

Top-4 limit experiments that are conducted in both global and local context of the model under *Ksim* similarity metric can be seen in Figure 7.21. Moreover the same experiments are also evaluated under *Osim* similarity metric, which can be found in Figure 7.22.

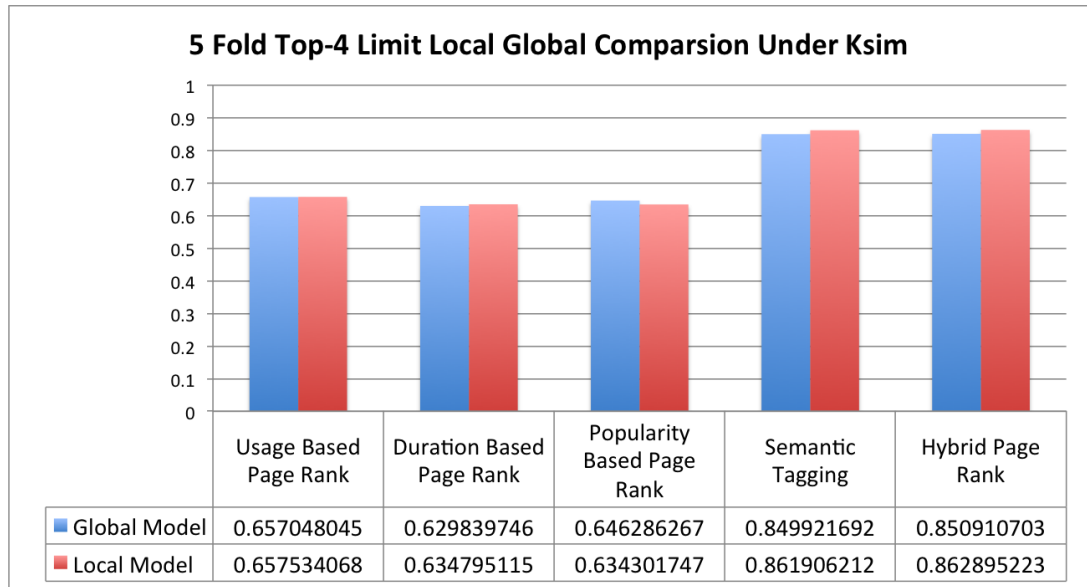


Figure 7.21: 5-Fold Validation with Top-4 Limit Under Ksim Similarity Metric

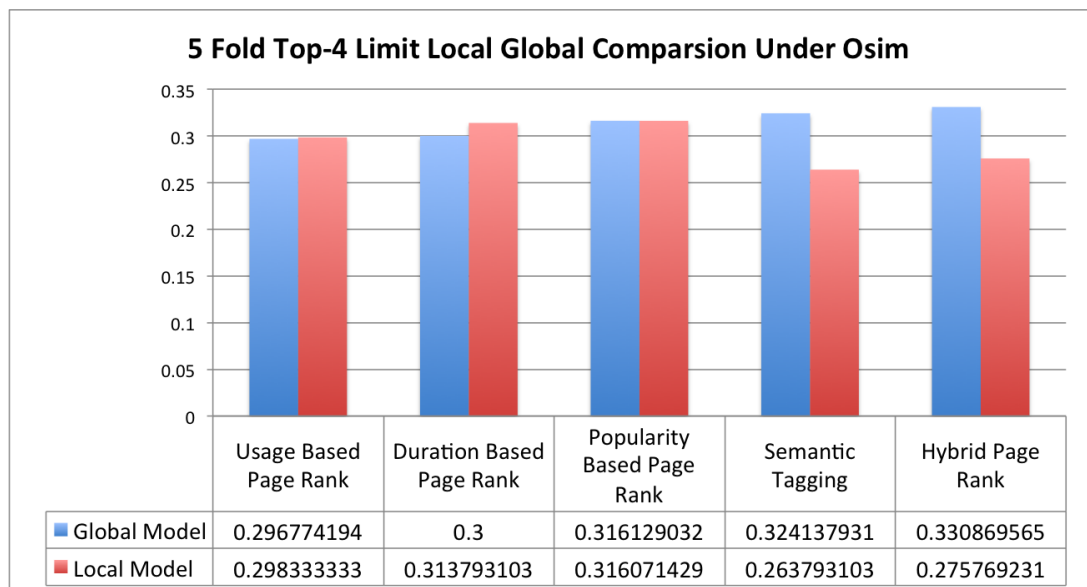


Figure 7.22: 5-Fold Validation with Top-4 Limit Under Osim Similarity Metric

When we analyze the results from 5-fold cross validation under *Ksim* for both global and

local models, Semantic Tagging and Hybrid Page Rank next page prediction models is approximately 30% more accurate than Usage Based Page Rank recommendations. Moreover in global modeling under *Osim*, both prediction models are more accurate than Usage Based Page Rank by 10%. However with local modeling, accuracy of both systems is decreased.

In order to investigate the effect of local model in comparison global model with top-4 next page predictions, we calculate the change percentage of local model to global model. In Figure 7.23, recommendations with limit value 2 are evaluated with two similarity metrics.

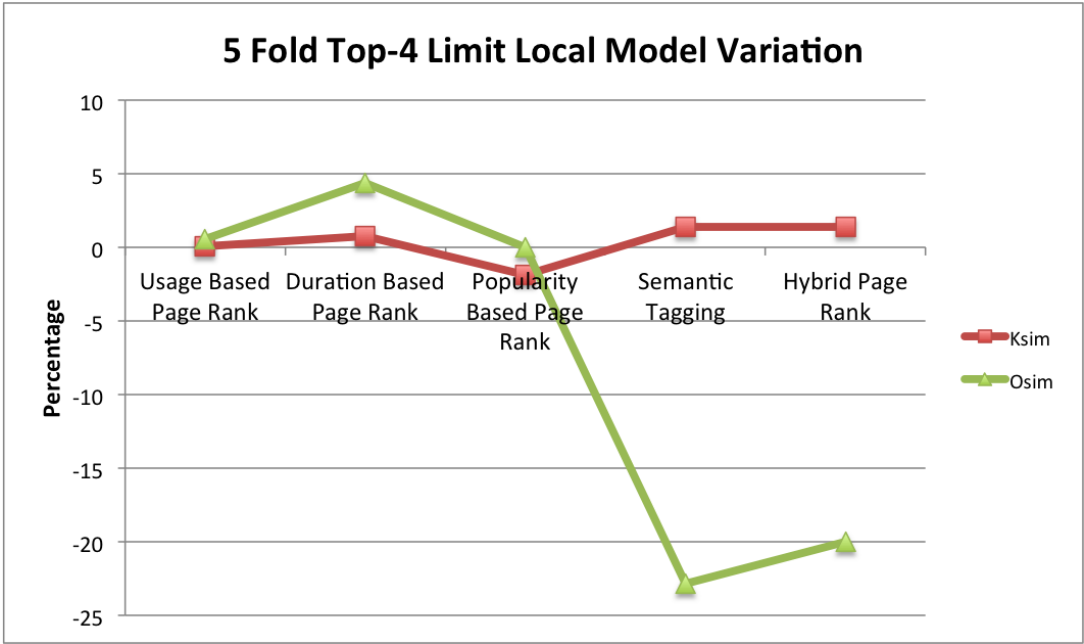


Figure 7.23: 5-Fold Validation with Top-4 Local Model Variation Percentage

### 7.2.3 5-Fold Cross Validation with Top-8 Limits

In top-8 limit of experiments, as being a maximum value, the general aim is to investigate the limit value behaviors of each next page prediction methodology. Like other experiments, they run in both global and local context of the model with top-8 limit under *Ksim* and *Osim* similarity metric. The results can be found in Figure 7.24 and Figure 7.25.

We run top-8 experiments in 5-fold cross validation and again, Semantic Tagging and Hybrid Page Rank methods improve Usage Based Page Rank by 30% *Ksim* similarity metric for both global and local models. On the other hand, under *Osim* similarity metric, which focuses

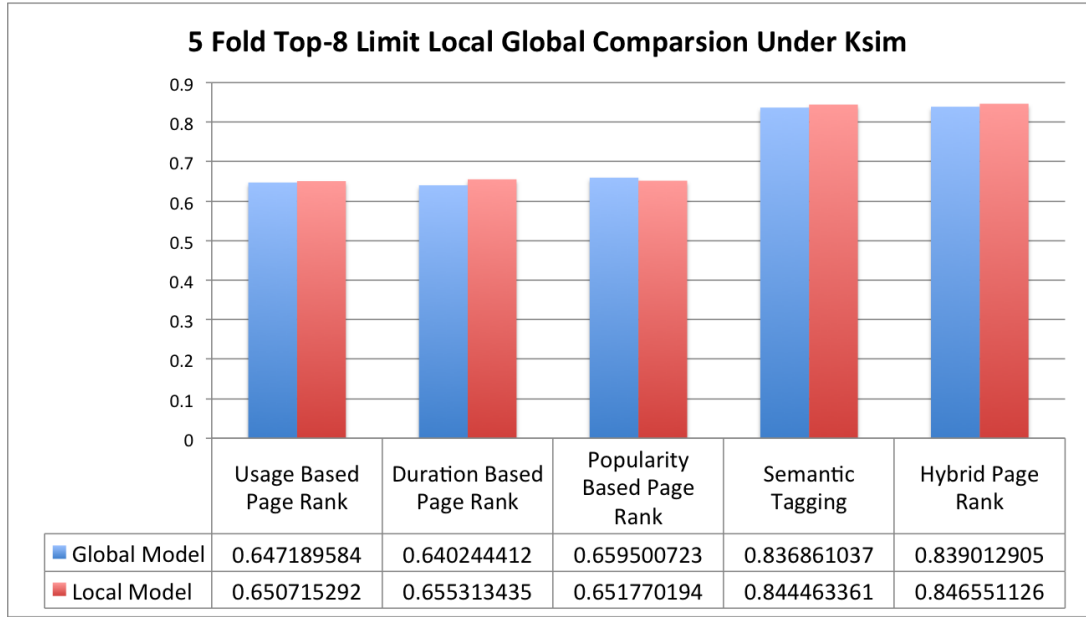


Figure 7.24: 5-Fold Validation with Top-8 Limit Under Ksim Similarity Metric

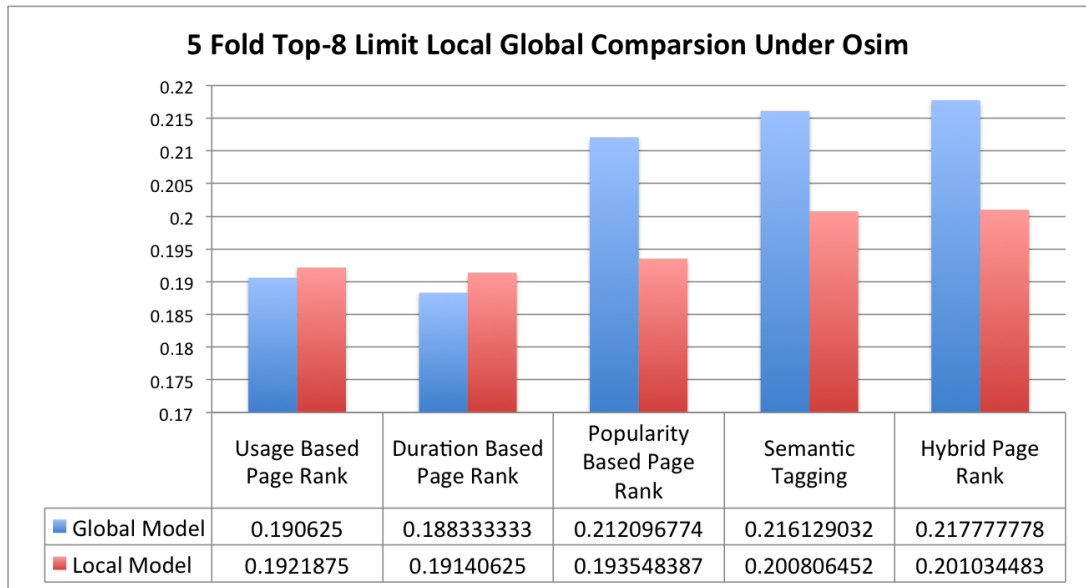


Figure 7.25: 5-Fold Validation with Top-8 Limit Under Osim Similarity Metric

only on the common elements in compared lists, experiment result shows an improvement on specified methods by 13% in global model and with local model improvement is nearly 5%.

In order to investigate the effect of local model in comparison global model with top-4 next page predictions, we calculate the change percentage of local model to global model. In

Figure 7.26, recommendations with limit value 2 are evaluated with two similarity metrics. From this results, it is obvious that preferring global model has an improvement over local model in Semantic Tagging and Hybrid Based Page Rank methods.

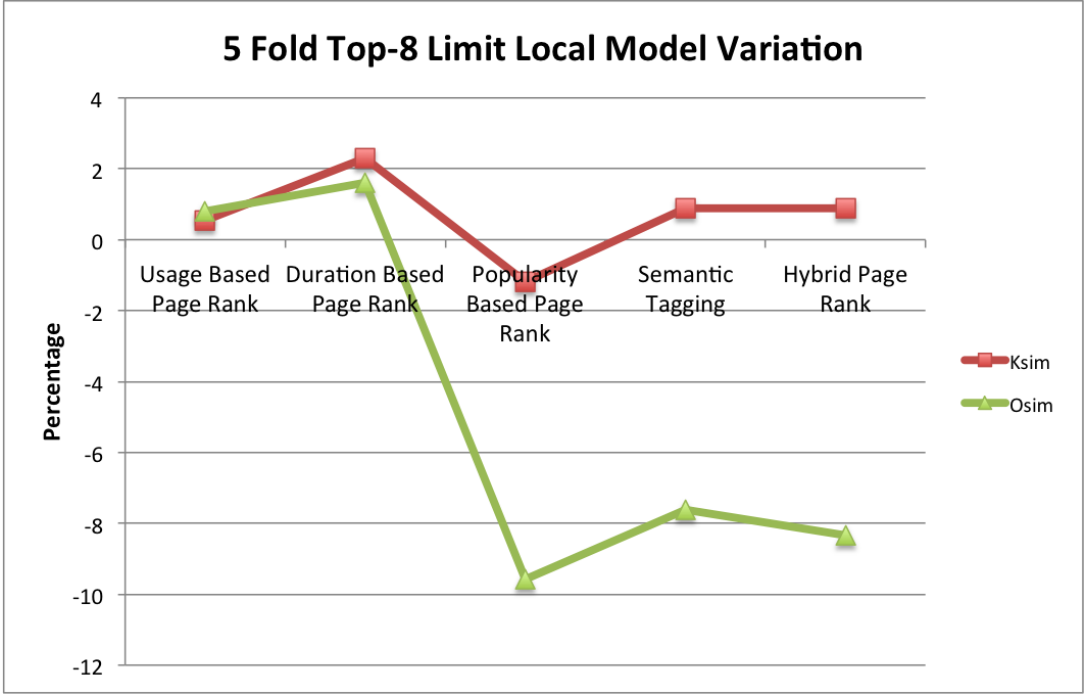


Figure 7.26: 5-Fold Validation with Top-8 Local Model Variation Percentage

#### 7.2.4 General Results

Although the accuracy of each methodology is increased by increasing the training and test data common values, there is still a positive effect of Semantic Tagging and Hybrid Based Page Rank methods on next page predictions. Under *Ksim* similarity metric, both of the methods in global and local models, they are 30% more accurate than Usage Based Page Rank. With *Osim* similarity, this effect is increased to 10% in global modeling. Moreover, it is observed that with local modeling, effectiveness of Semantic Tagging and Hybrid Based Page Rank are obviously decreased.

In Figure 7.27, Figure 7.28, Figure 7.29 and Figure 7.30 the general results of each next page prediction method with local and global models and top-n limit values can be seen as a summary.



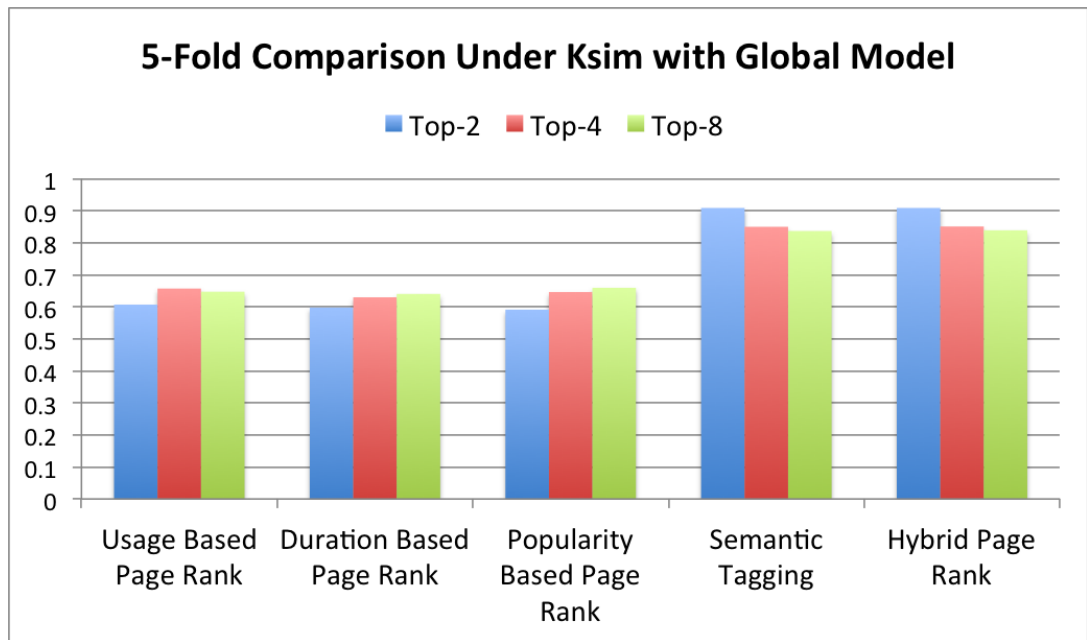


Figure 7.27: 5-Fold Ksim Comparison in Global Model

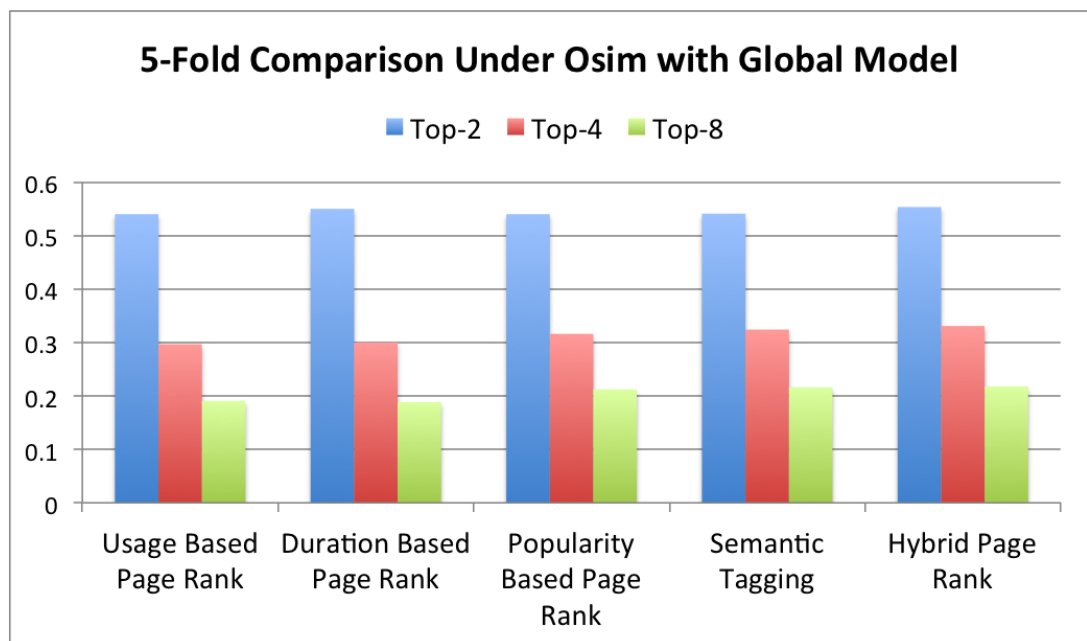


Figure 7.28: 5-Fold Osim Comparison in Global Model

In addition, the difference in variation of weights in Hybrid Page Rank model is also investigated through accuracy analysis. In the experiments we generate 9 different models and assign values starting with proportion of 0.1 to Semantic Tagging and eventually 0.9 to Pop-

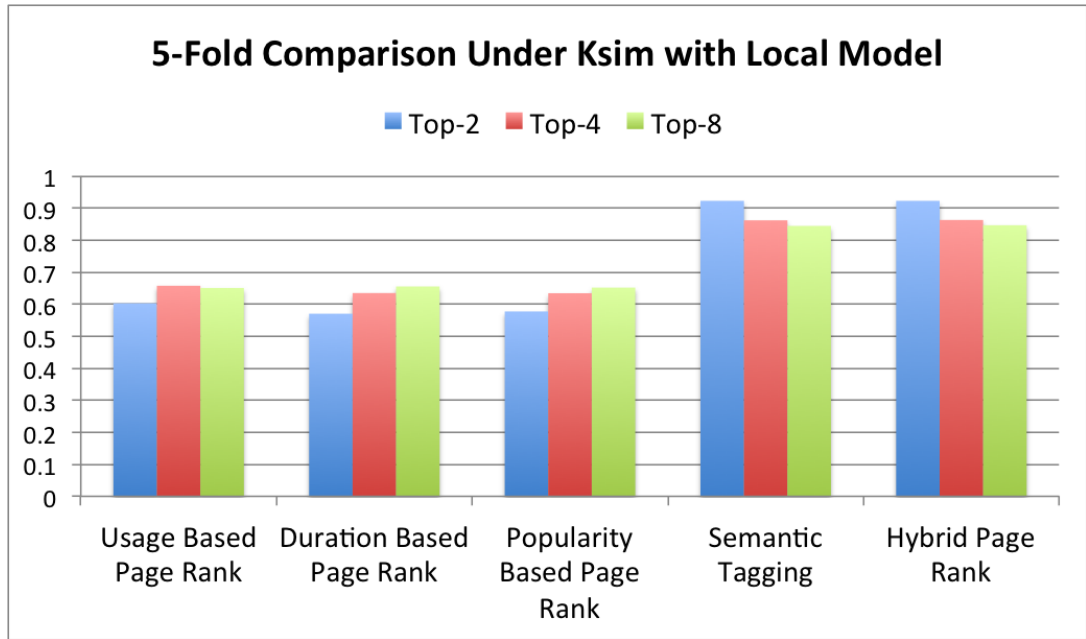


Figure 7.29: 5-Fold Ksim Comparison in Local Model

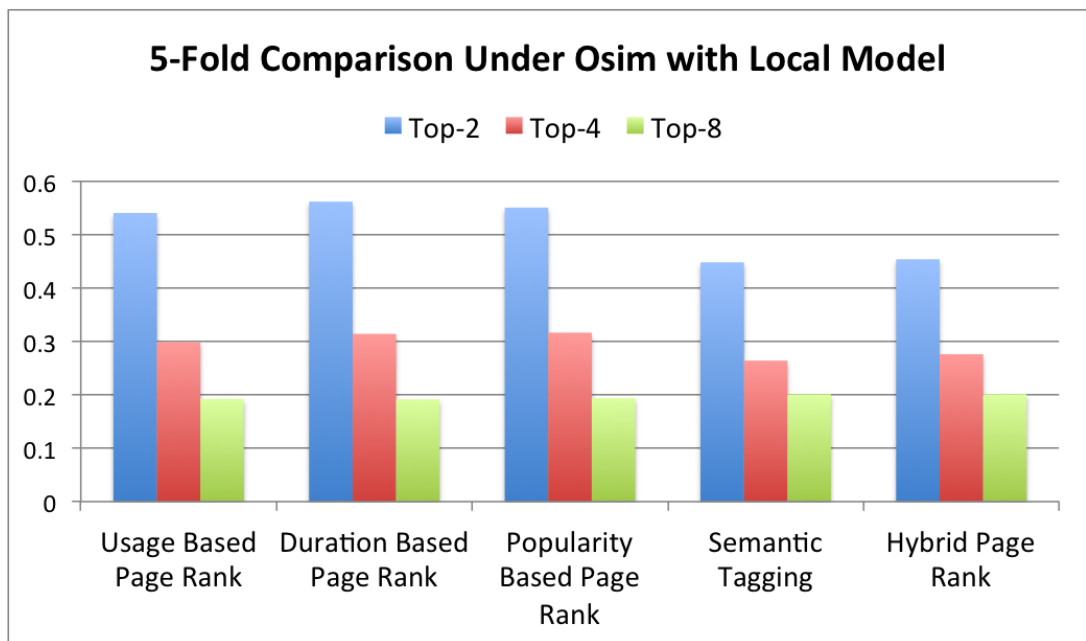


Figure 7.30: 5-Fold Osim Comparison in Local Model

ularity Based Page Rank. Then we iterate each run through increasing the effect of Semantic Tagging with 0.1 and eventually increasing Popularity Based Page Rank weight with 0.1. The results for each similarity metric and model can be found in Figure 7.31, Figure 7.32, Figure

7.33 and Figure 7.34.

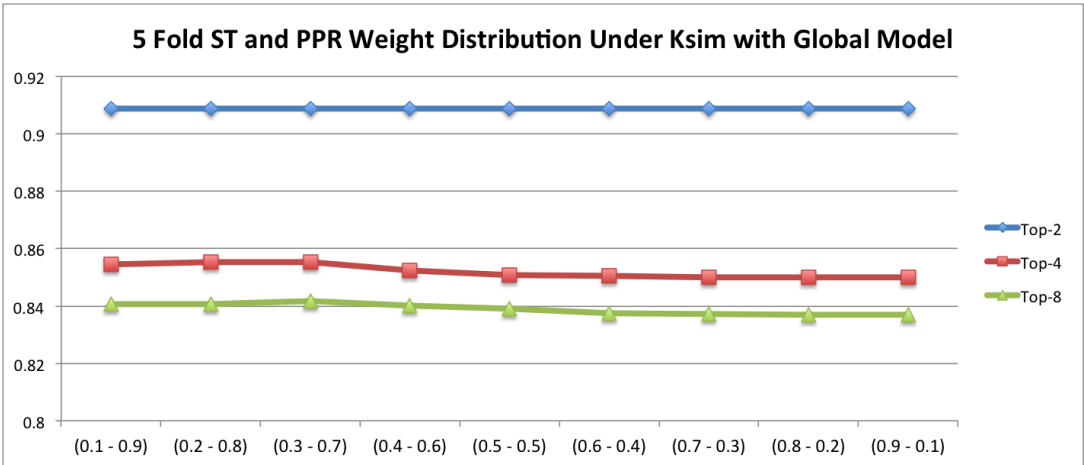


Figure 7.31: 5 Fold Validation ST and PPR Weight Effects on HPR under Ksim with Global Model

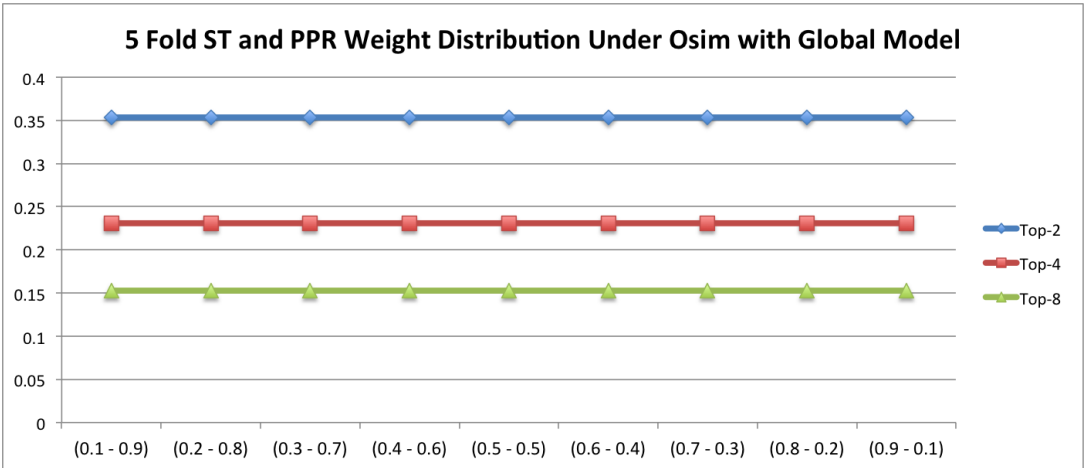


Figure 7.32: 5 Fold Validation ST and PPR Weight Effects on HPR under Osim with Global Model

Under *Ksim* similarity, (0.2 - 0.8) or (0.3 - 0.7) pairs can be preferred which is closer to Popularity Based Page Rank. On the other hand, under *Osim* similarity each weight distribution’s behavior is similar, since the common values of each next page prediction system is very closer to each other. In conclusion, pin hybrid model, (0.3 - 0.7) pair can be chosen due to the *Ksim* similarity metric effectiveness.

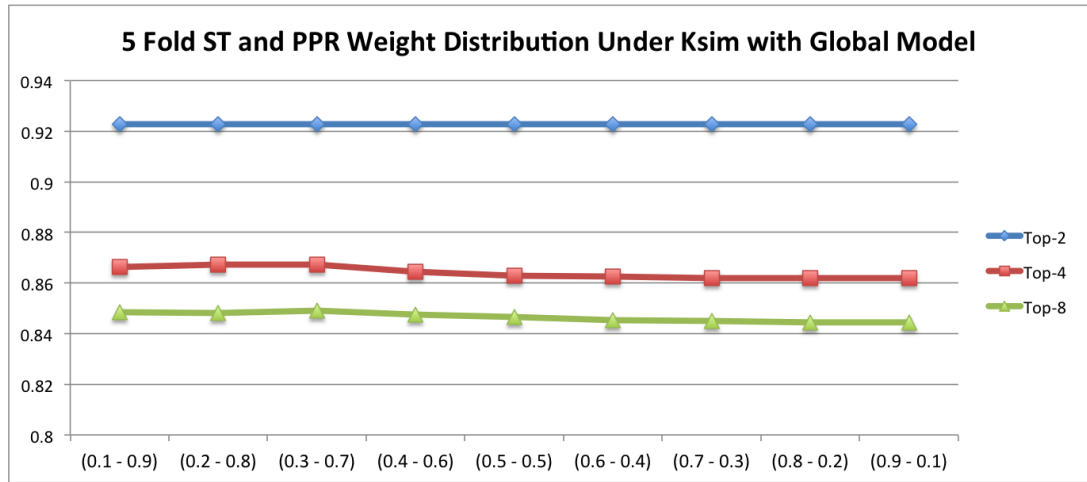


Figure 7.33: 5 Fold Validation ST and PPR Weight Effects on HPR under Ksim with Local Model

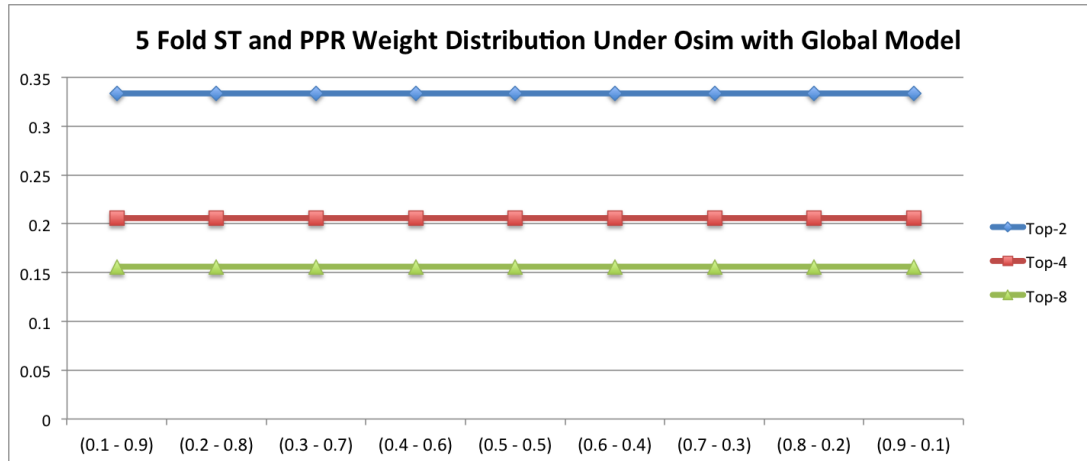


Figure 7.34: 5 Fold Validation ST and PPR Weight Effects on HPR under Osim with Local Model

### 7.3 10-Fold Cross Validation Experiments

Lastly we run our experiments with 10-fold cross validation and for this reason we divide data into ten parts and in each iteration we pick one of the not previously chosen for test data. Moreover in our setup, we make separate iterations with top-2, top-4 and top-8 recommendation comparisons that are measured by *Ksim* and *Osim* with global and local ranking methods. All five next page prediction model's (UPR, DPR, PPR, ST and HPR) accuracy under *Ksim* and *Osim* similarity metrics and local and global models are given at the rest of this section.

#### 7.3.1 10-Fold Cross Validation with Top-2 Limits

The experiments that are conducted in both global and local context of the model with top-2 limit under *Ksim* similarity metric can be seen in Figure 7.35. Moreover the same experiments are run and evaluate with *Osim* similarity metric, which can be found in Figure 7.36.

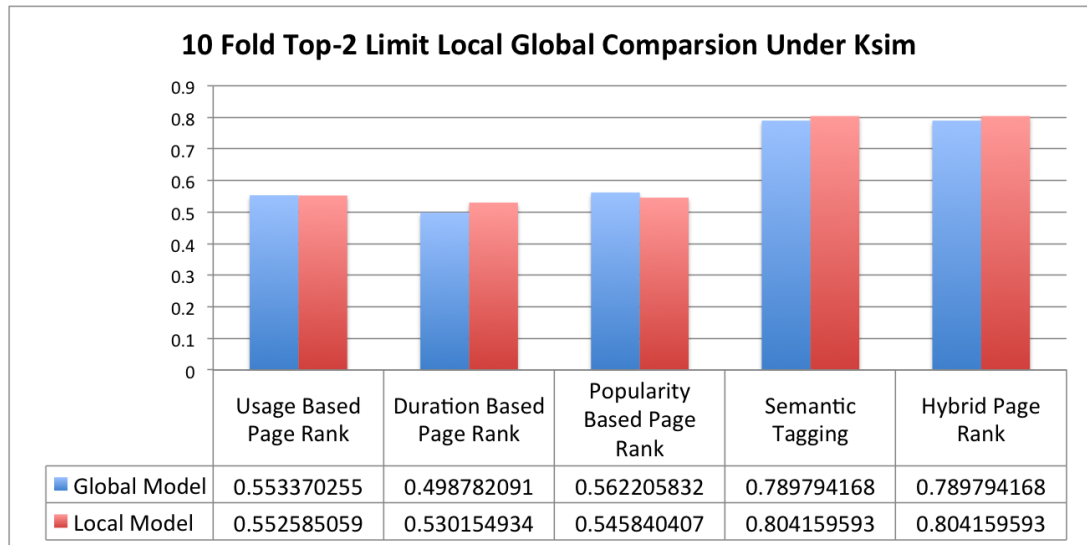


Figure 7.35: 10-Fold Validation with Top-2 Limit Under Ksim Similarity Metric

Under *Ksim* similarity metric, both Semantic Tagging and Hybrid Page Rank approaches improved the previous works with nearly 45% for both global and local model. However by measuring the *Osim* similarity, the effectiveness of the Semantic Tagging and Hybrid Page Rank becomes negative with 15 percentage. On the other hand Popularity Based Page Rank predictions are 5% effective than Usage Based Page Rank predictions in global model and in local model the identification between them is absolutely very close to each other. In conclu-

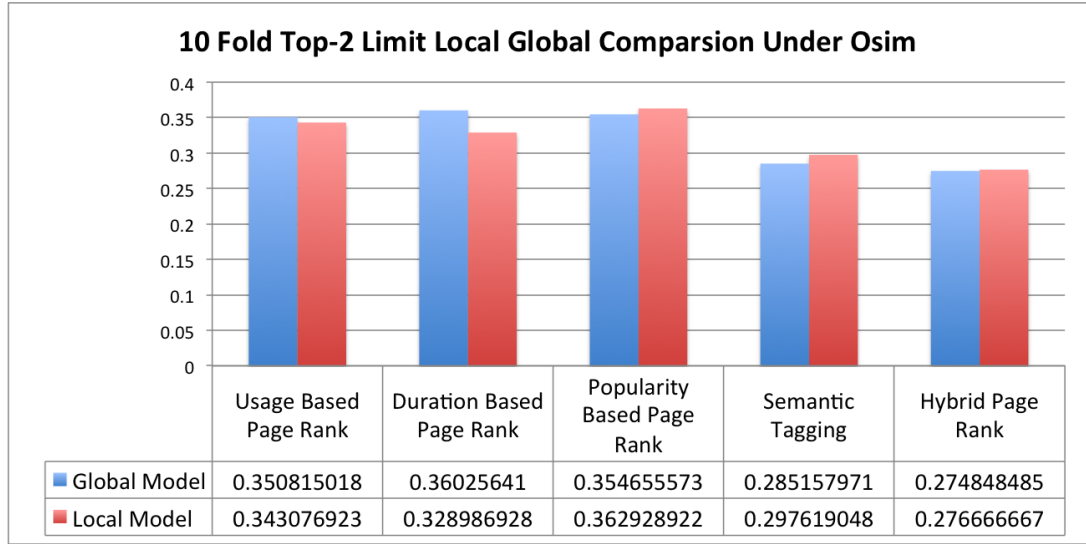


Figure 7.36: 10-Fold Validation with Top-2 Limit Under Osim Similarity Metric

sion, with 10 fold validation with top-2 limit recommendations, although Semantic Tagging and Hybrid Page Rank methodologies under *Ksim* similarity is effective, since the *Osim* metric shows that the real visiting values and number of predictions decreases comparison to other models. It should be point out that, for all prediction models, *Osim* similarity metric is under 0.4 value which shows that actually any methods are not accurate enough on this data set, produced from 10-fold. The same problem mentioned in 5-fold validation is higher in this situation. Hence, by comparing the common values of next page predictions and real visitings, all methods can be seen as identical. However, when we add the ordering factor to next page predictions by measuring *Ksim*, it can be obtained that ST and HPR predictions are more accurate than related works.

In order to investigate the effect of local model to global model while calculating next page predictions, we calculate the change percentage of local model to global model. In Figure 7.37, recommendations with limit value 2 are evaluated with two similarity metrics.

### 7.3.2 10-Fold Cross Validation with Top-4 Limits

Top-4 limit experiments that are conducted in both global and local context of the model under *Ksim* similarity metric can be seen in Figure 7.38. Moreover the same experiments are run and evaluate with *Osim* similarity metric, which can be found in Figure 7.39.

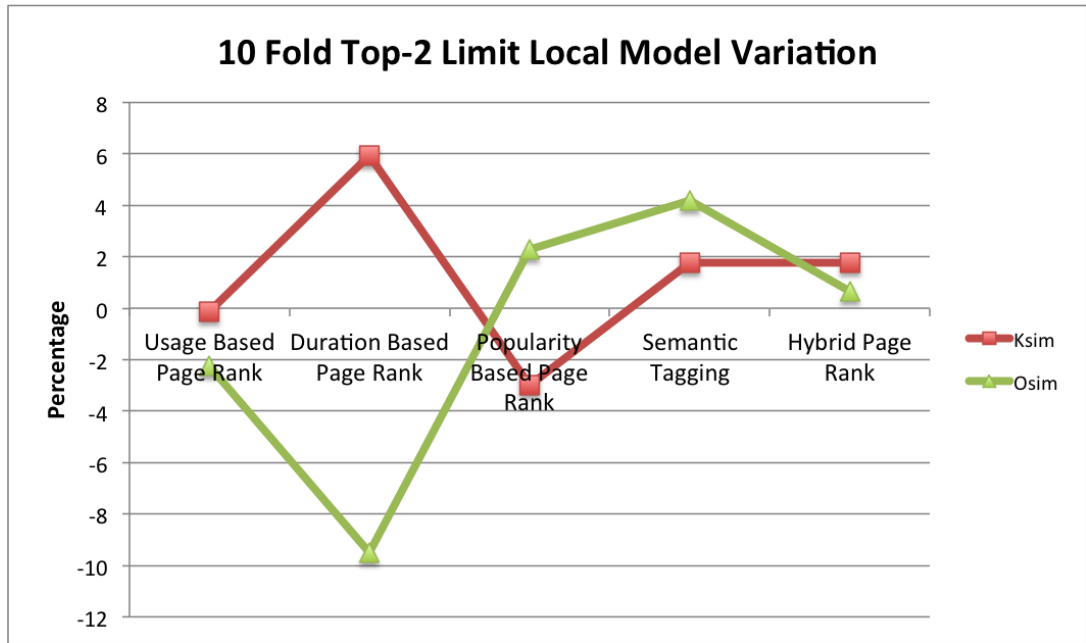


Figure 7.37: 10-Fold Validation with Top-2 Limit Local Model Variation Percentage

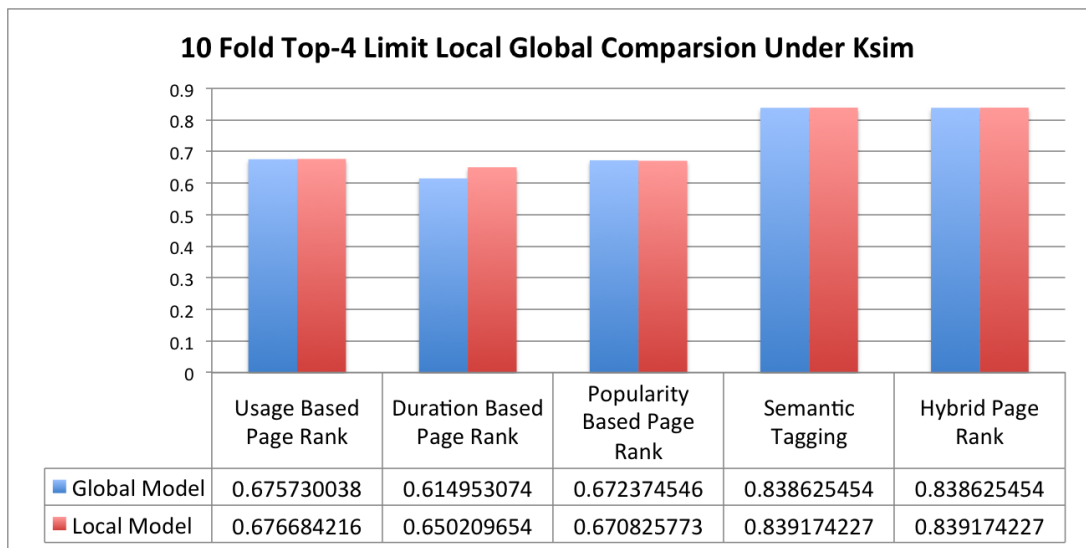


Figure 7.38: 10-Fold Validation with Top-4 Limit Under Ksim Similarity Metric

In Figure 7.38, it can be seen that under *Ksim* metric, Semantic Tagging and Hybrid Based Page Rank methods are 24% more accurate than Usage Based Page Rank in both local and global models. In Semantic Tagging predictions under *Osim*, in global model increase in accuracy is 7% and in local model the accuracy increase is nearly by 2%. At this point, again for Popularity Based Page Rank, the improvement is around 15% in comparison to

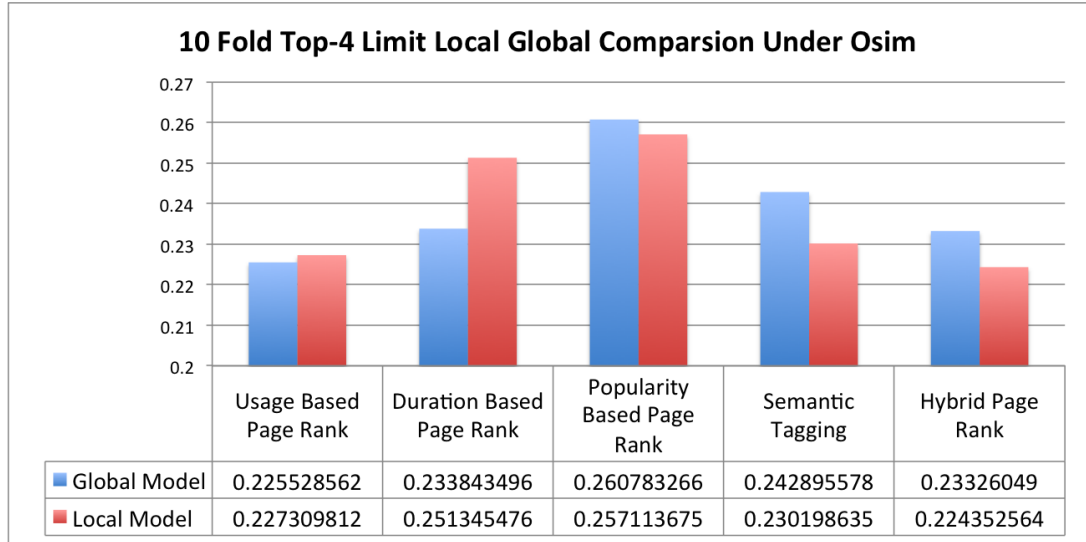


Figure 7.39: 10-Fold Validation with Top-4 Limit Under Osim Similarity Metric

Usage Based Page Rank. In addition to this, when measuring similarity with *Osim*, with all prediction methods the accuracy value is below 0.3, which shows actually all methods are identical and ineffective through the common number of real sets and recommendation sets based comparison.

In order to investigate the effect of local model to global model with top-4 next page predictions, we calculate the change percentage of local model to global model. In Figure 7.40, recommendations with limit value 2 are evaluated with two similarity metrics.

### 7.3.3 10-Fold Cross Validation with Top-8 Limits

In top-8 limit of experiments, as being a maximum value, the general aim is to investigate the limit value behaviors of each next page prediction methodology. Like other experiments, they run in both global and local context of the model with top-8 limit under *Ksim* and *Osim* similarity metric. The results can be found in Figure 7.41 and Figure 7.42.

In order to investigate the effect of local model to global model with top-4 next page predictions, we calculate the change percentage of local model to global model. In Figure 7.43, recommendations with limit value 2 are evaluated with two similarity metrics.



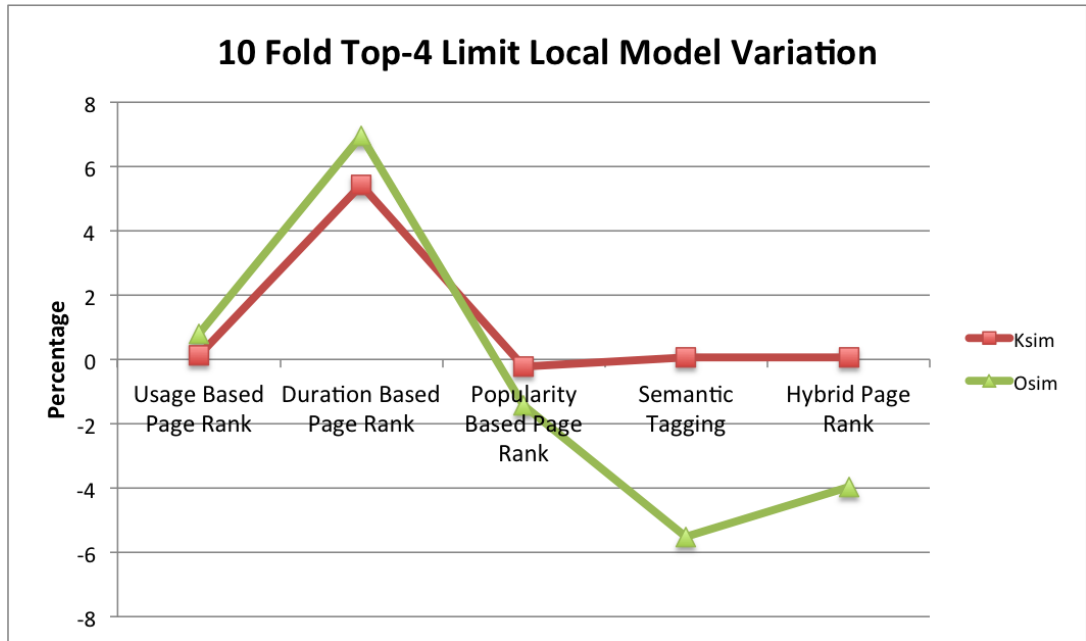


Figure 7.40: 10-Fold Validation with Top-4 Local Model Variation Percentage

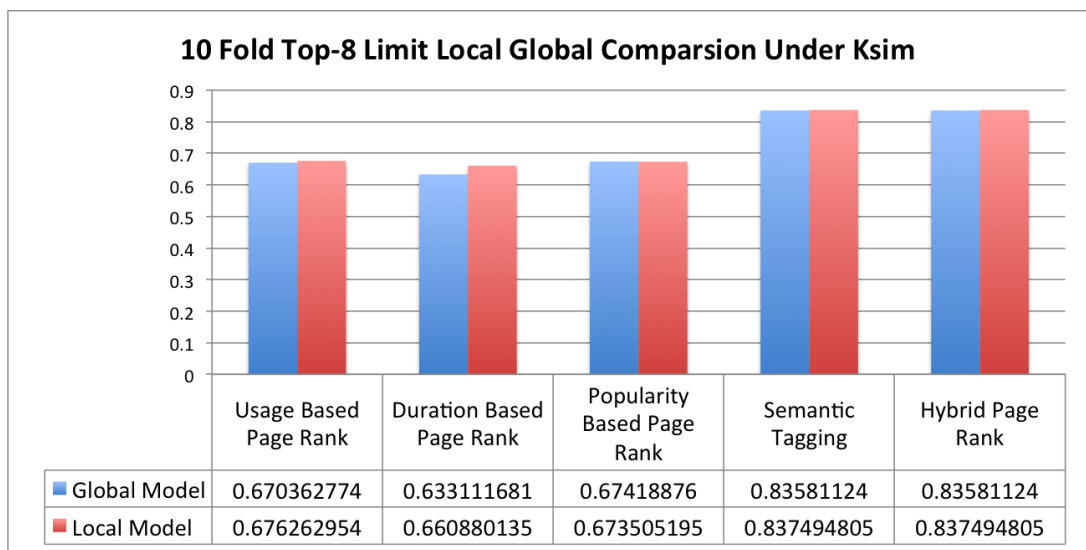


Figure 7.41: 10-Fold Validation with Top-8 Limit Under Ksim Similarity Metric

### 7.3.4 General Results

In 10-fold data partition, for all methods, finding common elements in real data with the training data set becomes a rare situation, since in our domain the diversity of pages is very high, training data does not cover most of the visits in test data, which makes all results below

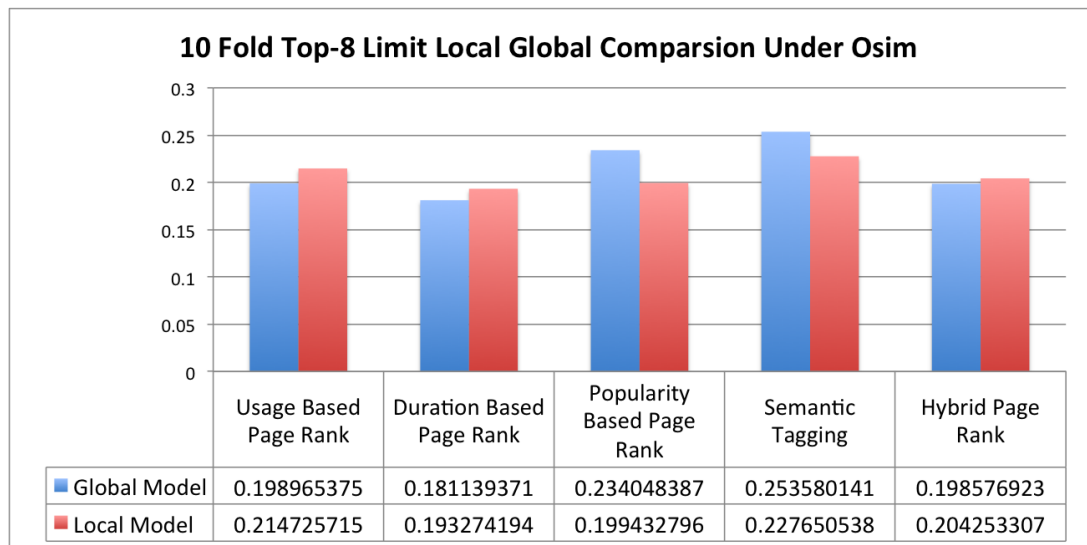


Figure 7.42: 10-Fold Validation with Top-8 Limit Under Osim Similarity Metric

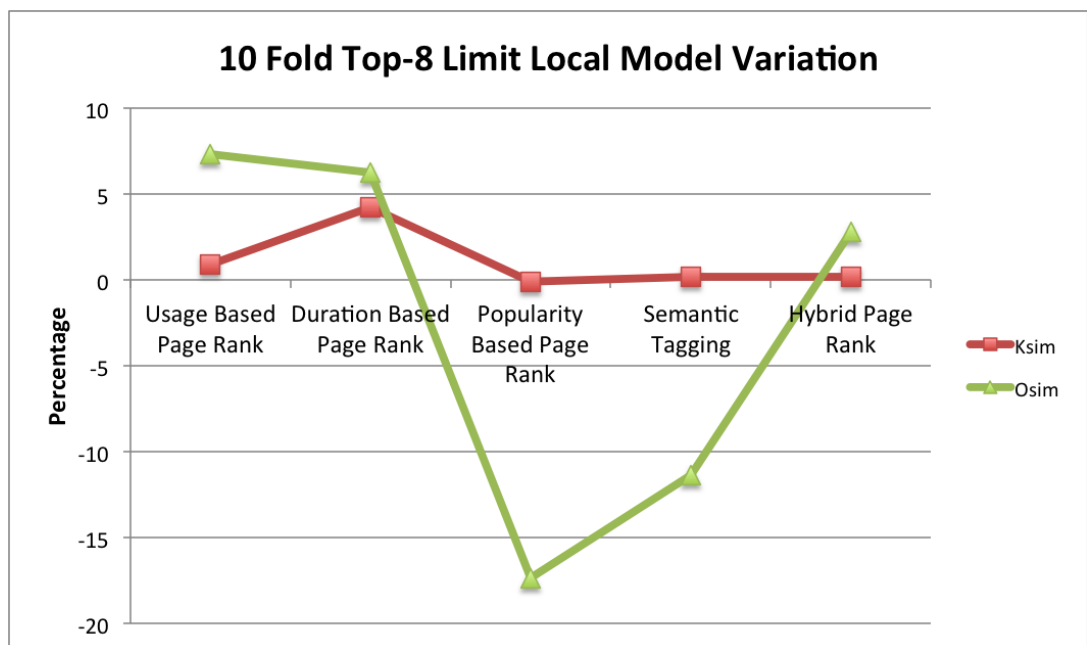


Figure 7.43: 10-Fold Validation with Top-8 Local Model Variation Percentage

0.4 with *Osim* similarity. On the other hand, with adding the effect of the order of the next page candidates with *Ksim* metric, it is observed that for both top-n limit, ST and HPR is effective than Usage Based Page Rank with nearly 35%. Moreover in this experiments it can be observed that using Popularity Based Page Rank can be preferred and generally sung global model produces more effective results.

In Figure 7.44, Figure 7.45, Figure 7.46 and Figure 7.47 the general results of each next page prediction method with local and global models and top-n limit values can be seen as a summary.

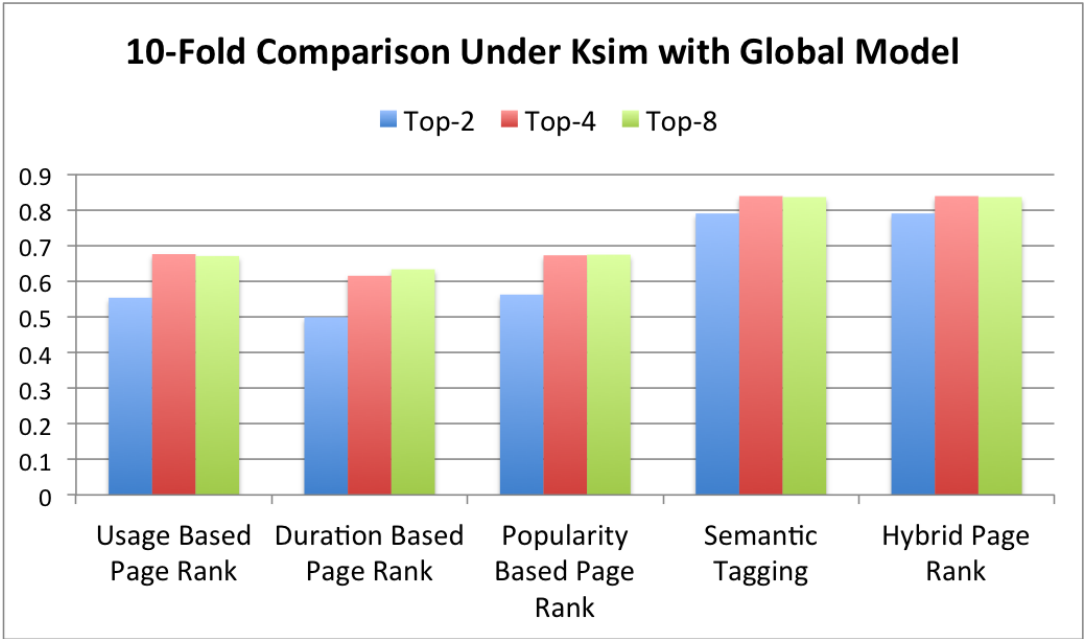


Figure 7.44: 10-Fold Ksim Comparison in Global Model

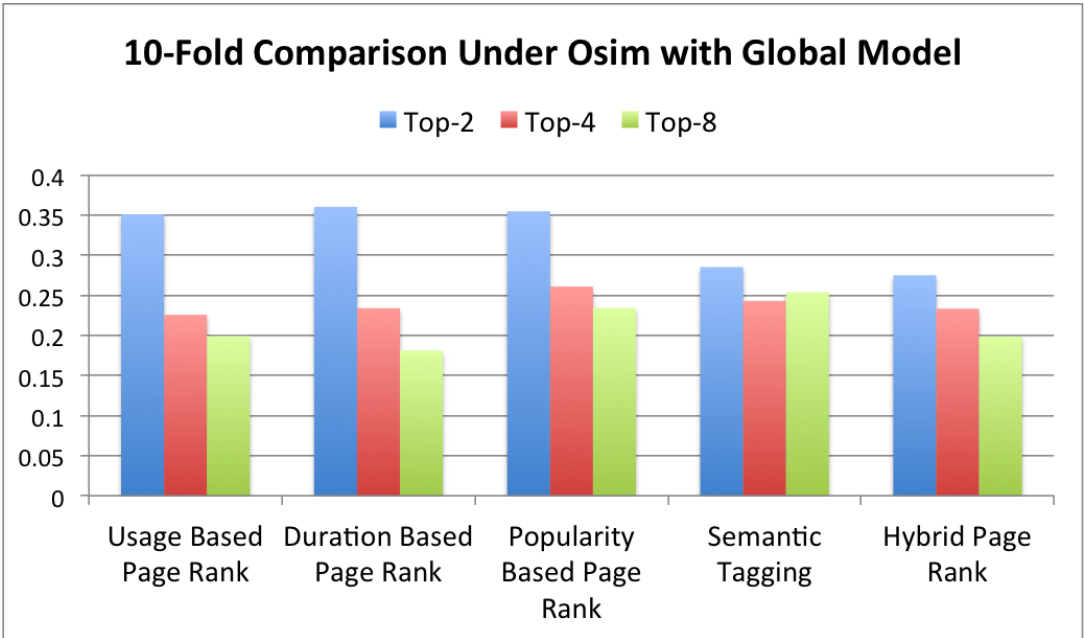


Figure 7.45: 10-Fold Osim Comparison in Global Model

In every different folded experiments, we make experiments for understanding the variation

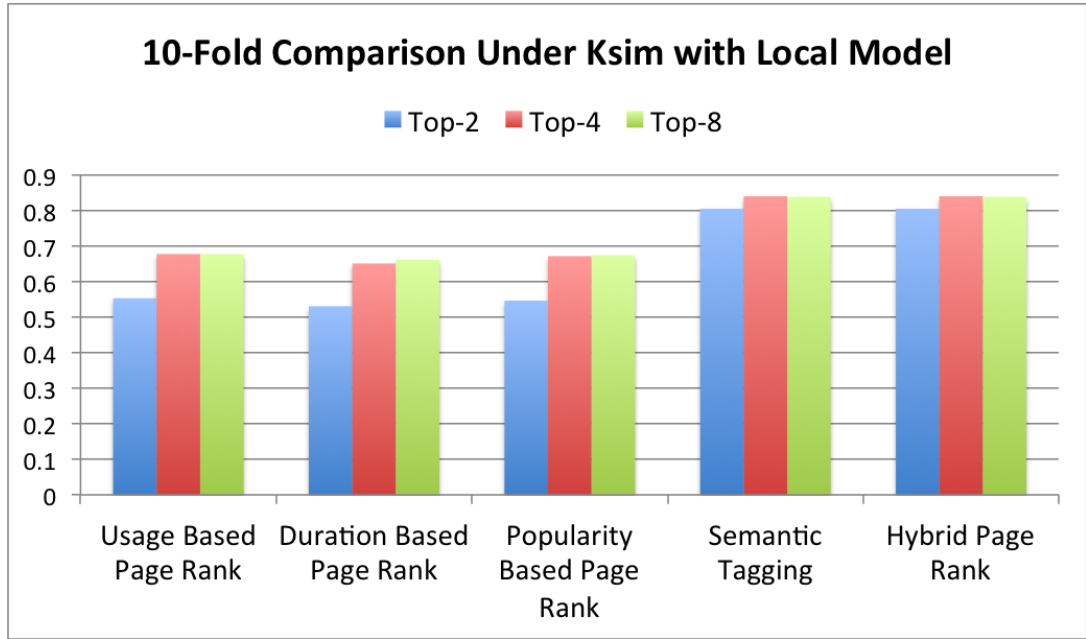


Figure 7.46: 10-Fold Ksim Comparison in Local Model

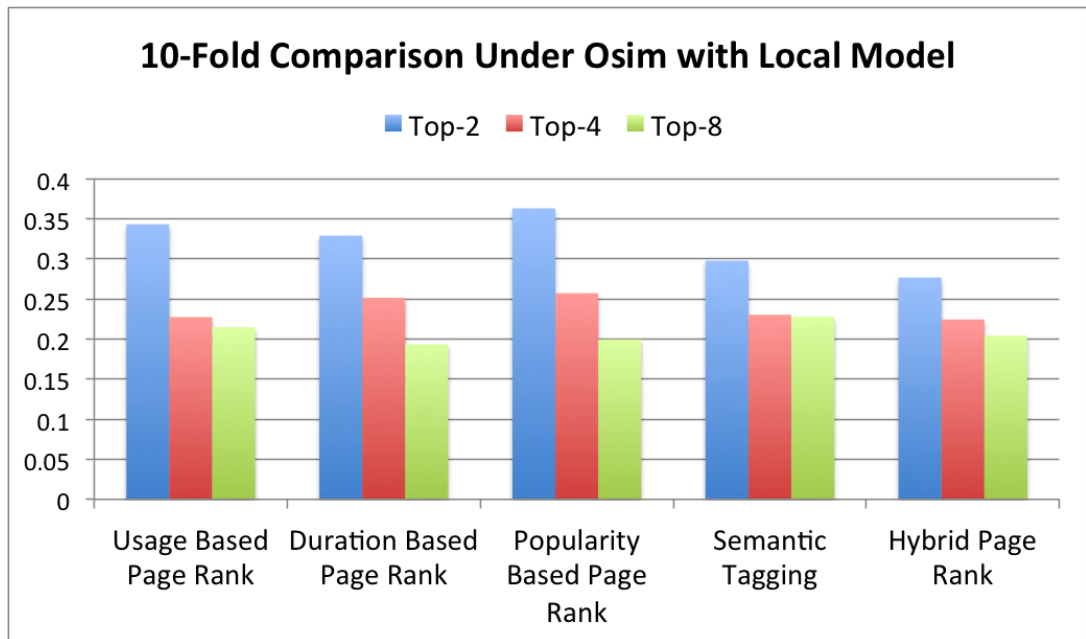


Figure 7.47: 10-Fold Osim Comparison in Local Model

of weights of each model to Hybrid Page Rank (HPR) model. We make experiments in 10-fold also. However, in this experiments it is observed that, for every calculation the results are so close to each other. This situation can be explained with two factors. In *Ksim* similarity,

the results do not change since the effect of Semantic Tagging (ST) is very high comparing to Popularity Based Page Rank (PPR). Since the effect of ST is so dominant, the HPR results always converges to ST. On the other hand, under *Osim* similarity, ST and PPR results are closer to each other and the identification casualty of that HPR values does not vary through the change of the weights of ST and PPR values.

#### 7.4 Evaluating Local and Global Modeling Effectiveness

For evaluating how local and global modeling affect the accuracy of next page predictions, we apply hypothesis t-test with global and local models. Hence in this test, we analyze the effect of modeling on next page prediction accuracy. For each fold values, we evaluate t-test results with confidence interval 99%. In our t-tests since we know that each value is the same measurement we use one tailed pairwise t-tests. We apply all of the method's results in each iteration for *Ksim* and *Osim* similarity measures.

Table 7.2: P-Value for Each Fold

	Top-2		Top-4		Top-8	
	Ksim	Osim	Ksim	Osim	Ksim	Osim
<b>3-Fold</b>	0.332	0.052	0.128	0.274	0.100	0.417
<b>5-Fold</b>	0.274	0.116	0.191	0.087	0.025	0.060
<b>10-Fold</b>	0.017	0.379	0.009	0.326	0.003	0.166

In these t-tests

- $h_0$  hypothesis is "There is no statistical significance between global and local modeling in accuracy of similarity measures."
- $h_1$  hypothesis is "There is a statistical significance between global and local modeling in accuracy of similarity measures."
- $\alpha = 0.01$

If the  $p\text{-value} < \alpha$  then we can say that the local and global modeling difference is statistically significant in calculating the accuracy of similarity measures in 99% confidence interval. In our calculations none of the results is smaller than the  $\alpha$  value which means the  $h_0$  hypothesis

is accepted with 99% confidence. On the other hand, if we accept  $\alpha$  value as 0.05 then for 10-fold cross validation the modeling can vary the accuracy of similarity measures.

### 7.5 Deciding Best k-value on Cross Validation

As a rule of thumb, using k value as 10 is the common behavior in cross validation. However in [32] it is also emphasized that this k value may be different with selected data's several attributes. In addition to this, in their work they show that with increasing level of k, the standard deviation is also increasing between folds, which produces unbiased results with that subject. In our experiments for each fold with *Ksim* and *Osim* separately, we calculate the standard deviation between each fold's accuracy results. Since in previous section, we show that with confidence interval of 99% the modeling is not statistically significant, we choose global models in our experiments.

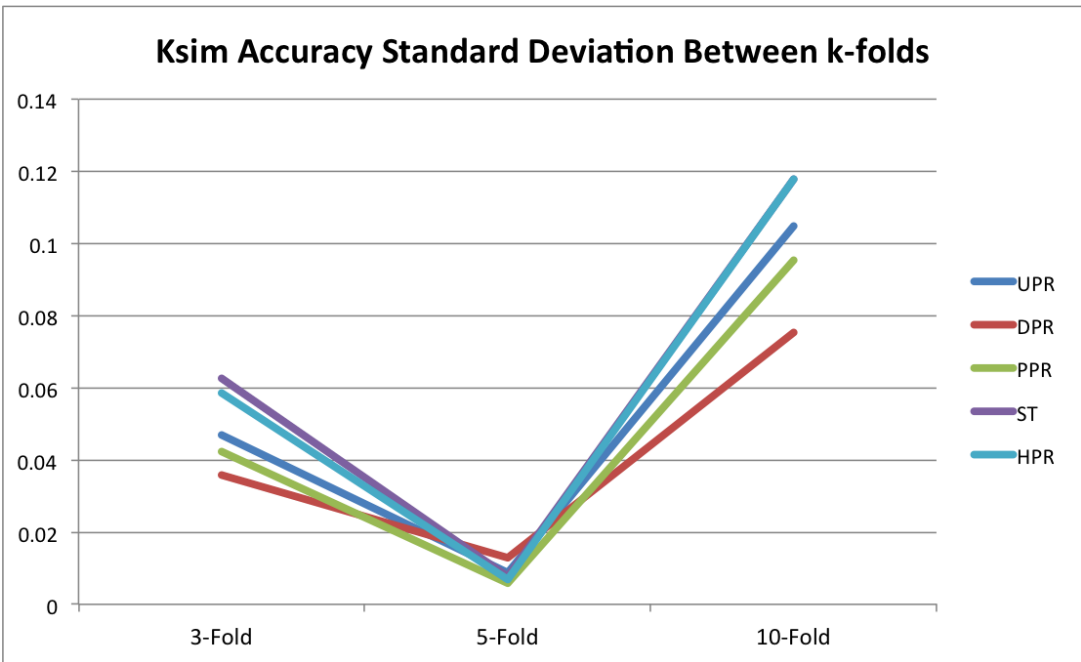


Figure 7.48: Ksim Standard Deviation in Global Model

In Figure 7.48 and Figure 7.49, it is obvious that with  $k = 5$ , we observe the *elbow* (details can be found in [35]) of the standard deviation values of k folds and moreover we see that this observation is valid for all the methods in the graphs. For this reason, it can be said that k value above 5 is not accurate with the data that we worked on.

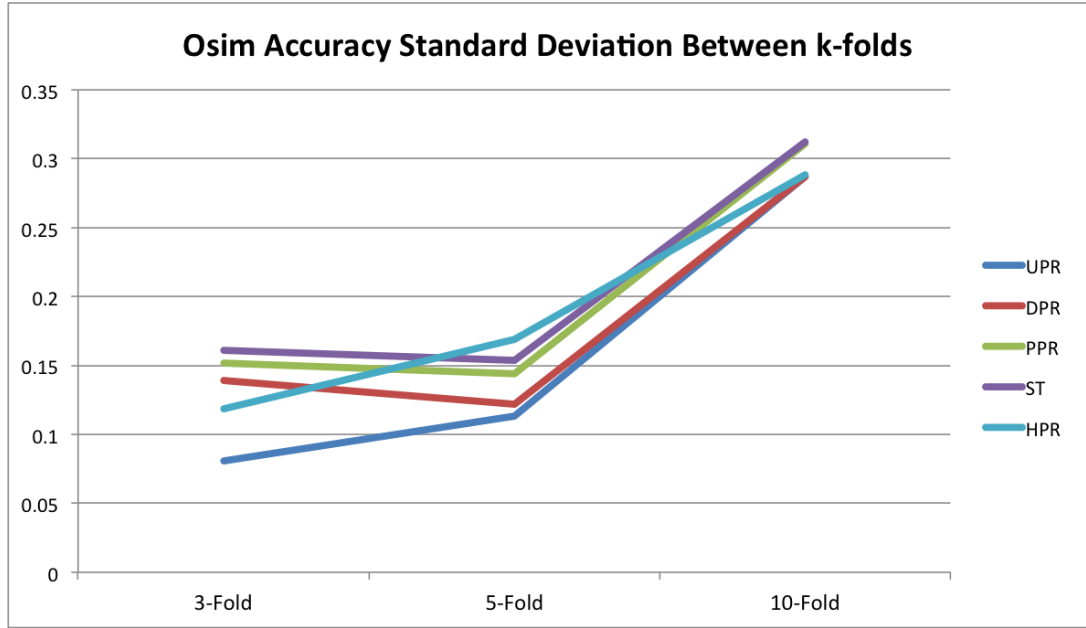


Figure 7.49: Osim Standard Deviation in Global Model

## 7.6 Precision and Recall Values at Top-8

As additional experiments, we calculate precision and recall values of each recommendation in each fold with top-8 limit values. Since we accept that the difference between global and local model is not significant, we make our experiments with only global modeling. In our calculations, instead of picking up a limited value of recommendation candidate (i.e 15), we apply all test fold to recommendation evaluation and in every recommendation, we record the precision and recall value and calculate the average value of all folds for each of the methods. In Figure 7.50, 7.51 and 7.52 the precision recall values of 3-fold, 5-fold and 10-fold is given, respectively. In each figure all methods are shown as points in scatterplot diagrams with precision in Y-axis and recall is in X-axis.

It is obvious that for each of the folds, Semantic Tagging (ST) approach is better than other methods. In order to identify the accuracy of each model, we support our calculations with  $F_1$  values of each method and fold, which are shown in Figure 7.53.

As a result, when we compare the accuracy of ST with UPR in terms of precision and recall values of each recommendation, we observe that in 3-fold cross validation, there is an improvement nearly by 48%, in 5-fold cross validation this values is nearly 38% and in 10-fold

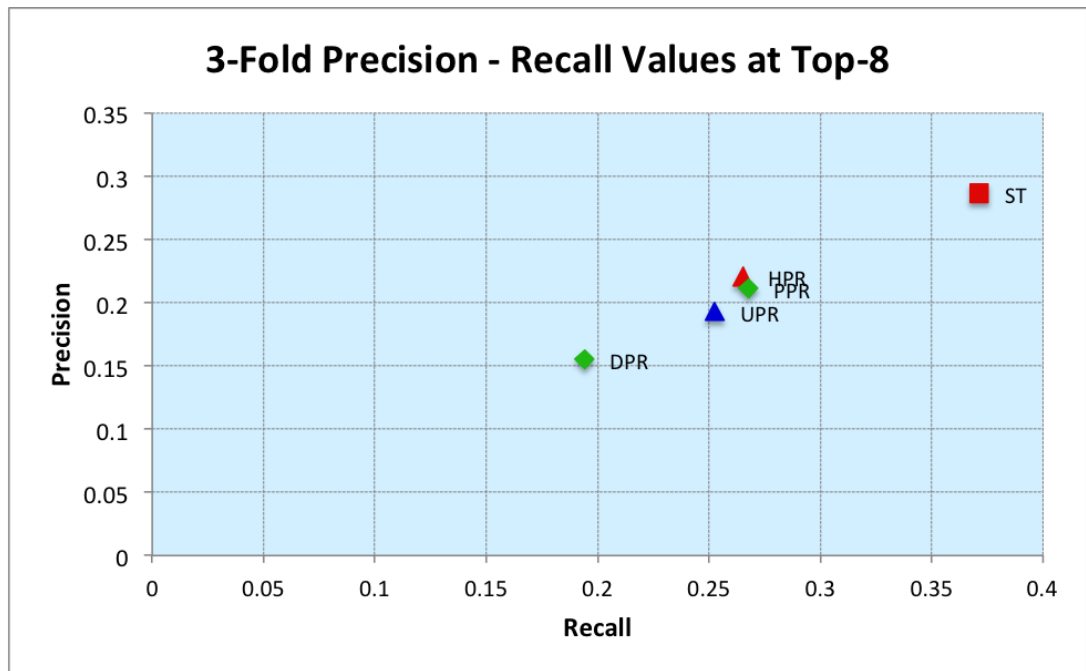


Figure 7.50: 3 Fold Precision Recall Values

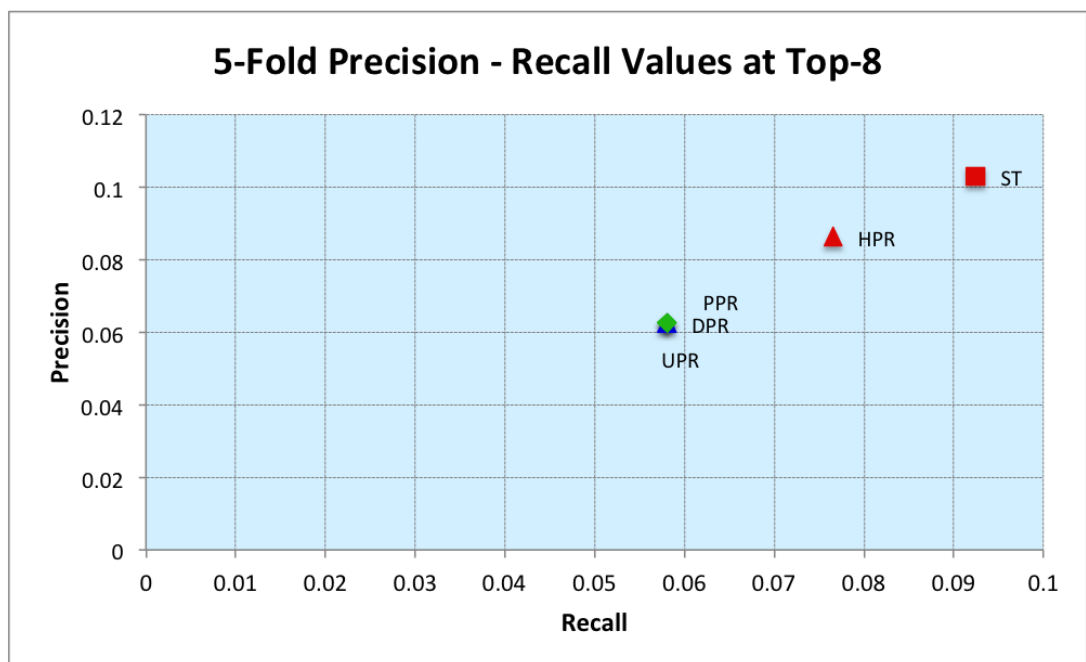


Figure 7.51: 5 Fold Precision Recall Values

cross validation this value is 7%.



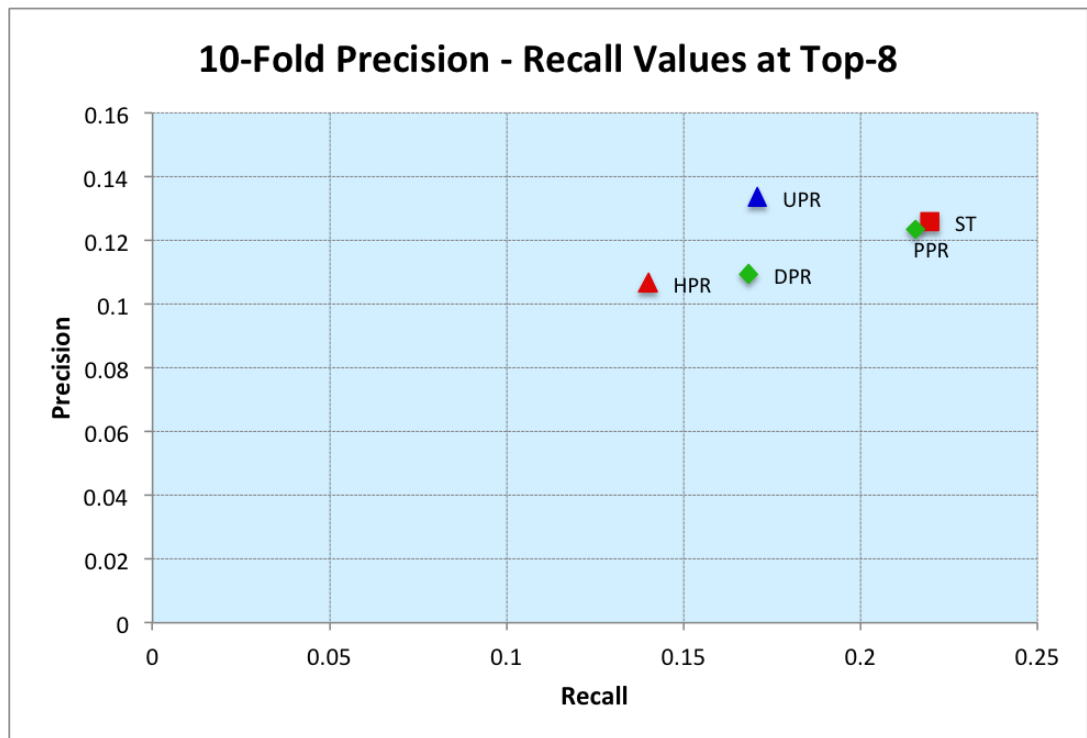


Figure 7.52: 10 Fold Precision Recall Values

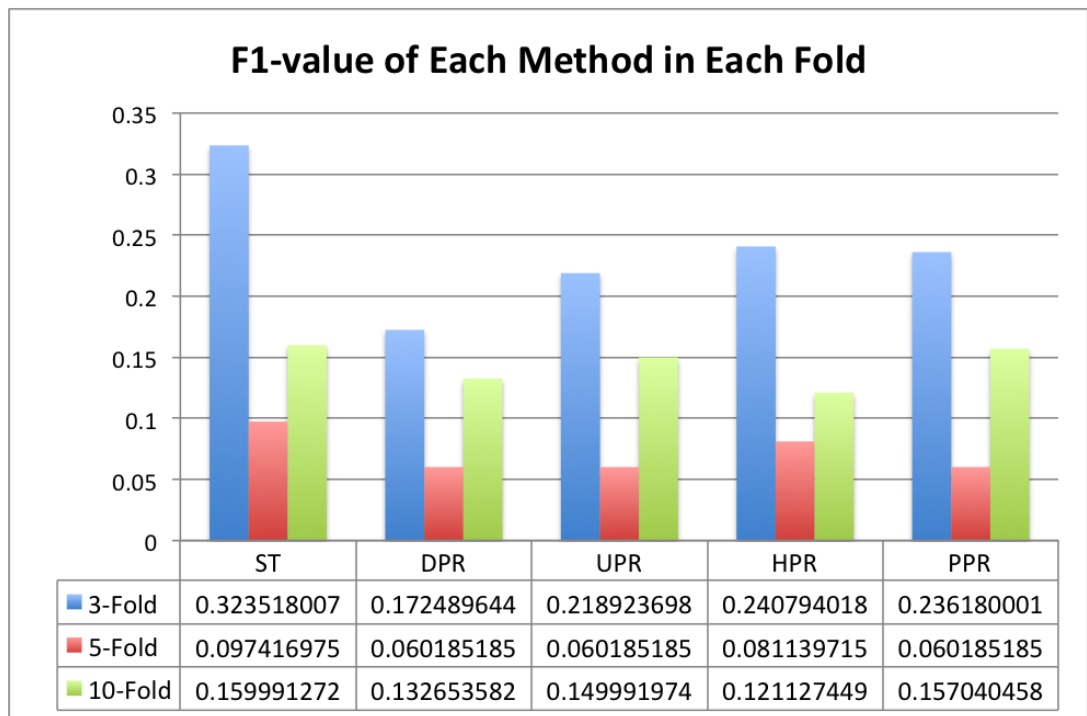


Figure 7.53:  $F_1$  Values of Each Method and Fold

## **CHAPTER 8**

### **CONCLUSION**

Page rank algorithms are commonly used for both next page prediction and web searching. There are several page ranking modeling methods that focus basically on frequency of pages and transition frequencies. In addition to this, duration of page visits retrieved from transitions can be considered as well [3]. However the duration of page, which can be directly related to page size, is not modeled for page ranking algorithm. For example if the user is waiting for the download of a long page including large objects such as images, obviously it would take more time than a page which includes really small amount of data. Although just the size information of page cannot produce information for popularity of a page, the proportion of duration and size can produce information for popularity of pages. We model this situation as Duration Based Page Rank (DPR), which concerns duration vs. size proportion. In addition, we model another hybrid approach that concerns both duration size proportion and frequency of pages and transitions which is called Popularity Based Page Rank (PPR) algorithm.

Commonly in all page rank algorithms web usage mining is applied in order to get information about user navigations. In our experiments, we observe that Duration Based Page Rank (DPR) and Popularity Based Page Rank (PPR) algorithms are improved the previous works [3, 6], although a little. For modeling a next page prediction system which can have more accurate results than our previous works, we decide to add our method web content mining.

In web content mining, there are various techniques for pulling out the information related to user navigation in the navigated pages. These information can be obtained from web page's content or just from the URL. Since in our data, the URL includes several data that can be used as attributes of pages and we do not prefer to add this work's scope information retrieval techniques, we choose to use web URLs for retrieving semantic information of web sites.

Semantic Tagging (ST) approach is used to categorize web pages with information only related to its URLs. With our novel semantic similarity calculation, we calculate each page pair's similarity for finding the most similar page for next page prediction with last visited page information. However in our calculation, if the pair of pages semantic similarity is equal, which is very common especially with pages that similarity value with lower values, as a support tool we use PPR results in each next page candidates to sort them in the recommendation lists.

Hybrid Page Rank (HPR) approach, we model Semantic Tagging (ST) and Popularity Based Page Rank (PPR) with equal powers for calculating the new value as HPR, and in next page prediction opposite of ST, we sort HPR values of candidate pages in the recommendation lists.

Moreover in Hybrid Page Rank experiments generally we apply 0.5 proportion from each method for constructing hybrid page rank. In spite of this generalization, we make some extra experiments for finding the best proportion between ST and PPR values. In all of the experiments we observe that in 0.3 - 0.7 proportion (ST with 0.3 and 0.7 with PPR) pair, accuracy becomes its maximum value, and we observe that the accuracy is very rarely more than Semantic Tagging (ST) method. In addition to this, we observe that this method never returns results least than ST or PPR alone. As a result, it can be obtained that this method do not produce a far more better results comparing to Semantic Tagging which uses PPR as a support tool.

With 3-fold and 5-fold we observe that Semantic Tagging (ST) has an improvement at least 25% comparing to UPR with *Ksim* and *Osim* similarity metrics. Moreover we observe that especially with *Ksim* measurement, this improvement is increased to 35% values, which considers both common elements and the *order* of the pages in recommendation list and real visiting list.

In recommendation systems, recommending more than two pages may be a rare situation for next page predictions. In that point of view, *Ksim* similarity can be seen as more important metric than *Osim* especially in short recommendation lists. On the other hand, marketing web sites like "Amazon, e-bay, etc." recommends several products (as pages) which can lower the importance of ordering in recommendations a little. Therefore for deciding the limit values of recommendations, one of the most important factor is the deciding the aim of the related web site. If this is a marketing web site, higher values of limits are preferred.

In [6], in order to shorten the respond time of calculating page rank values, which is a recursive and complex algorithm, a web site's graph is modeled as a synopsis of it as they called local model. In our calculations we record intermediate steps in database in order to decrease the calculation time of the page rank for both local and global models. Since we have a chance to prefer the best modeling in calculating the page rank values, we calculate accuracy of each five next page prediction method (UPR, DPR, PPR, ST, HPR) with both global and local with each cross validation experiments. In these experiments we observe that the effect of local and global modeling in page rank calculation is very weak, however in order to support our decision with statistical methods, we apply t-tests for each top-n value with each cross validation. Then we observe that the effect of modeling on page rank accuracy is not significant in 99% confidence.

Furthermore we apply each of our experiment cross validation in order to decrease the *chance* factor of our results with swapping test data with all the elements in the data set. In cross validation, as a general idea of choosing 10 fold is not the best fit for all types of data [32]. In order to see the effect of each cross validation, we calculate standard deviation of each fold's accuracy results with each other. In these experiments we observe that after  $k = 5$ , the standard deviation of accuracy between fold values is increased dramatically. Therefore it can be said that with this data set, increasing the  $k$  value after 5 decreases the homogeneity of iterations inside cross validation.

In the semantic tagging process we pick the most frequent pages (frequency threshold is 10) and we tag them each related concepts manually. As a future work, automatic tagging can be developed in order to decrease the effort of manually tagging. Furthermore these experiments can be applied into another domain with this automated process. On the other hand, in semantic tagging process, some association rules can be defined for next page predictions.

In next page predictions it is always a hard situation to find a solution to the *cold start* case of the web usage mining process' natural. As a remedy of this, supporting this type of situations with web structural information can be considered.

Next page prediction is a promising and useful area and it becomes a need in all of the web sites, especially in marketing. However all the web sites need them, it is not appropriate to use one kind of solution to all of these web sites. The first condition of modeling the best next page prediction system is to make an investigating about the domain of web site and

web server logs of it. If the web server logs are not appropriate, for satisfying the modeling, a web site plugin should be developed, though it gets several risks with itself. For finding the suitable next page prediction is a hard situation, although all web sites want to gain that power to navigate their users into web pages that the users want to visit in next step. Our work aims to answer some questions with several parameters related to this problem.

## REFERENCES

- [1] M. M. Group, "Internet world stats - usage and population statistic." <http://www.internetworldstats.com/stats.htm/>, 2011. Last Access: 2011 October 14.
- [2] M. D. Kunder, "World wide web size - daily estimated size of the world wide web." <http://www.worldwidewebsite.com/>, 2011. Last Access: 2011 November 23.
- [3] Y. Z. Guo, K. Ramamohanarao, and L. Park, "Personalized pagerank for web page prediction based on access time-length and frequency," in *Web Intelligence, IEEE/WIC/ACM International Conference on*, pp. 687–690, November 2007.
- [4] H. Liu and V. Kešelj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data Knowl. Eng.*, vol. 61, pp. 304–330, May 2007.
- [5] N. Duhan, A. Sharma, and K. Bhatia, "Page ranking algorithms: A survey," in *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pp. 1530–1537, March 2009.
- [6] M. Eirinaki and M. Vazirgiannis, "Usage-based pagerank for web personalization," in *Data Mining, Fifth IEEE International Conference on*, p. 8 pp., nov. 2005.
- [7] S. Gunduz and M. T. Ozsu, "A web page prediction model based on click-stream tree representation of user behavior," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03*, (New York, NY, USA), pp. 535–540, 2003.
- [8] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM Trans. Internet Technol.*, vol. 3, pp. 1–27, February 2003.
- [9] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from web usage data," in *Proceedings of the 3rd international workshop on Web information and data management, WIDM '01*, (New York, NY, USA), pp. 9–15, 2001.
- [10] M. K. N. Wai-Ki Ching, *Markov Chains: Models, Algorithms and Applications*. Springer, 2005.
- [11] J. Borges and M. Levene, "Evaluating variable-length markov chain models for analysis of user web navigation sessions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, pp. 441–452, April 2007.
- [12] M. Deshpande and G. Karypis, "Selective markov models for predicting web page accesses," *ACM Trans. Internet Technol.*, vol. 4, pp. 163–184, May 2004.

- [13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
- [14] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, "Web path recommendations based on page ranking and markov models," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, (New York, NY, USA), pp. 2–9, ACM, 2005.
- [15] S. Madria, S. Bhowmick, W. Ng, and E. Lim, "Research issues in web data mining," in *Data Warehousing and Knowledge Discovery* (M. Mohania and A. Tjoa, eds.), vol. 1676 of *Lecture Notes in Computer Science*, pp. 805–805, Springer Berlin / Heidelberg, 1999.
- [16] S. Paulakis, C. Lampos, M. Eirinaki, and M. Vazirgiannis, "Sewep: A web mining system supporting semantic personalization," in *Knowledge Discovery in Databases: PKDD 2004* (J. F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, eds.), vol. 3202 of *Lecture Notes in Computer Science*, pp. 552–554, Springer Berlin / Heidelberg, 2004.
- [17] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu, "Integrating web usage and content mining for more effective personalization," in *Electronic Commerce and Web Technologies* (K. Bauknecht, S. Madria, and G. Pernul, eds.), vol. 1875 of *Lecture Notes in Computer Science*, pp. 165–176, Springer Berlin / Heidelberg, 2000.
- [18] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez, "Conceptual user tracking," in *Advances in Web Intelligence* (E. Menasalvas, J. Segovia, and P. Szczepaniak, eds.), vol. 2663 of *Lecture Notes in Computer Science*, pp. 955–955, Springer Berlin / Heidelberg, 2003.
- [19] Y.-F. B. W. Min Song, "Handbook of research on text and web mining technologies," in *Handbook of Research on Text and Web Mining Technologies*, vol. 1, pp. 386–401, IGI Global, 2008.
- [20] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, "Low-complexity fuzzy relational clustering algorithms for web mining," *Fuzzy Systems, IEEE Transactions on*, vol. 9, pp. 595–607, aug 2001.
- [21] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the world wide web," in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, pp. 558–567, nov 1997.
- [22] W. Xing and A. Ghorbani, "Weighted pagerank algorithm," in *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pp. 305–314, May 2004.
- [23] B. Berendt, A. Hotho, and G. Stumme, "Towards semantic web mining," in *In International Semantic Web Conference (ISWC)*, pp. 264–278, Springer, 2002.
- [24] T. Haveliwala, "Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, pp. 784–796, July-Aug. 2003.
- [25] G. Inc., "Google search engine." <http://www.google.com/>, 2011. Last Access: 2011 October 05.

- [26] W. Community, "Wikipedia the free encyclopedia." <http://en.wikipedia.org/wiki/PageRank/>, 2011. Last Access: 2011 December 12.
- [27] R. Dutta, A. Kundu, R. Dattagupta, and D. Mukhopadhyay, "An approach to web page prediction using markov model and web page ranking," *Journal of Convergence Information Technology*, vol. 4, no. 4, pp. 61–67, 2009.
- [28] X. Qi and B. D. Davison, "Web page classification: Features and algorithms," *ACM Comput. Surv.*, vol. 41, pp. 12:1–12:31, February 2009.
- [29] R. Kosala and H. Blockeel, "Web mining research: a survey," *SIGKDD Explor. Newsl.*, vol. 2, pp. 1–15, June 2000.
- [30] M.-Y. Kan, "Web page classification without the web page," in *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, WWW Alt. '04, (New York, NY, USA), pp. 262–263, ACM, 2004.
- [31] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: bounds for k-fold and progressive cross-validation," in *Proceedings of the twelfth annual conference on Computational learning theory*, COLT '99, (New York, NY, USA), pp. 203–208, ACM, 1999.
- [32] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pp. 1137–1143, 1995.
- [33] D. Anguita, S. Ridella, and F. Riviuccio, "K-fold generalization capability assessment for support vector classifiers," in *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 2, pp. 855 – 858 vol. 2, july-4 aug. 2005.
- [34] N. Jardine and C. van Rijsbergen, "The use of hierarchic clustering in information retrieval," *Information Storage and Retrieval*, vol. 7, no. 5, pp. 217 – 240, 1971.
- [35] D. J. KETCHEN and C. L. SHOOK, "The application of cluster analysis in strategic management research: An analysis and critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.



## APPENDIX A

### WEB LOG'S CAPTURED CONCEPTS and WEB URLS

In our experimental evaluations, we use METU's web server logs from 29/May/2010 to 18/Feb/2011. After pruning pages and capturing concepts related to these page's URLs, we map each URL to concepts. In Table A.1, concepts constituting the three levels of concept hierarchy can be found. In addition to this, in Table A.2, complete list of URL-concept mapping is given.

Table A.1: Three Level Concepts

Level Number	Concept Name
Level 1	Course
Level 1	Document
Level 1	Feed
Level 1	Grad
Level 1	Introduction
Level 1	Lecturer
Level 1	Library
Level 1	Misc
Level 1	Research
Level 1	Seminar
Level 1	Student Page
Level 1	Undergrad
Level 2	akilic
Level 2	alan
Level 2	Algorithm

Table A.1: (continued)

Level 2	All
Level 2	alpaslan
Level 2	Alumni
Level 2	Android
Level 2	aykut
Level 2	aysegul
Level 2	aysun
Level 2	birturk
Level 2	bozsahin
Level 2	bozyigit
Level 2	C++
Level 2	cagatay
Level 2	ceng 111
Level 2	ceng 140
Level 2	ceng 230
Level 2	ceng 232
Level 2	ceng 242
Level 2	ceng 280
Level 2	ceng 300
Level 2	ceng 302
Level 2	ceng 303
Level 2	ceng 334
Level 2	ceng 336
Level 2	ceng 350
Level 2	ceng 351
Level 2	ceng 352
Level 2	ceng 382
Level 2	ceng 436
Level 2	ceng 443
Level 2	ceng 444

Table A.1: (continued)

Level 2	ceng 462
Level 2	ceng 463
Level 2	ceng 465
Level 2	ceng 466
Level 2	ceng 469
Level 2	ceng 476
Level 2	ceng 477
Level 2	ceng 483
Level 2	ceng 490
Level 2	ceng 520
Level 2	ceng 536
Level 2	ceng 556
Level 2	ceng 562
Level 2	ceng 563
Level 2	ceng 564
Level 2	ceng 567
Level 2	ceng 568
Level 2	ceng 574
Level 2	ceng 584
Level 2	ceng 701
Level 2	ceng 705
Level 2	ceng 707
Level 2	ceng 713
Level 2	ceng 714
Level 2	ceng 732
Level 2	ceng 734
Level 2	cosar
Level 2	cuneyt
Level 2	deniz
Level 2	Discrete Math

Table A.1: (continued)

Level 2	dogru
Level 2	erdas
Level 2	erkut
Level 2	erman
Level 2	erol
Level 2	faculty
Level 2	Football
Level 2	ftitrek
Level 2	genc
Level 2	gokcen
Level 2	gokdeniz
Level 2	gtumuklu
Level 2	gulen
Level 2	Intern
Level 2	isler
Level 2	ismet
Level 2	karagoz
Level 2	kasim
Level 2	ketenci
Level 2	levent
Level 2	Music
Level 2	nafiz
Level 2	nebil
Level 2	News
Level 2	nihan
Level 2	oguztuzn
Level 2	onur
Level 2	orald
Level 2	otopcu
Level 2	Plugin

Table A.1: (continued)

Level 2	polat
Level 2	ruken
Level 2	sciftci
Level 2	se 548
Level 2	se 705
Level 2	selma
Level 2	sener
Level 2	sercan
Level 2	sertan
Level 2	sibel
Level 2	skalkan
Level 2	tcan
Level 2	toroslu
Level 2	ucoluk
Level 2	volkan
Level 2	vural
Level 2	yalabik
Level 2	yazici
Level 3	genc
Level 3	gtumuklu
Level 3	isler
Level 3	Java
Level 3	karagoz
Level 3	News
Level 3	nihan
Level 3	sibel
Level 3	tcan

Table A.2: Whole Data Set's Captured Concepts

Page URL	1st Level Concept	2nd Level Concept	3rd Level Concept
/_ export/raw/index	Misc	-	-
/_ export/xhtmll/index	Misc	-	-
/_ media/course/ceng111/sinem-demirci.jpeg?w=178&rev=	Course	ceng 111	-
/_ media/people/faculty/alpaslan/index.bib?rev=	Lecturer	alpaslan	-
/_ vti_ inf.html	Misc	-	-
/ % 7E dogru	Lecturer	dogru	-
/ % 7E e114068/guestbook6/gb.php	Student Page	-	-
/ % 7E e1195288/Java% 20Programming% 20Unleashed.pdf	Student Page	-	-
/ % 7E e120353/HoughTransform/FP_ analys.pdf	Student Page	-	-
/ % 7E e1402668/hw/index.php	Student Page	-	-
/ % 7E e1402668/hw/index.php	Student Page	-	-
/ % 7E e1402668/hw/index.php?do=add_ form&page=1	Student Page	-	-
/ % 7E e1416056/hw4/guestbook/guestbook.php	Student Page	-	-
/ % 7E e1416056/hw4/guestbook/guestbook.php	Student Page	-	-
/ % 7E e1448786/Evanescence% 20-% 2006% 20-% 20Understanding.MP3	Student Page	Music	-
/ % 7E genc/334/Ch_ 11_ Vista.ppt	Lecturer	genc	-
/ % 7E isler/page0001.html	Lecturer	isler	-

Table A.2: (continued)

/% 7Eisler/page0002.html	Lecturer	isler	-
/% 7Eisler/page0004.html	Lecturer	isler	-
/% 7Enihan/ceng302	Course	ceng 302	-
/% 7enihan/ceng302/dbms.ppt	Course	ceng 302	-
/% 7enihan/ceng302/dbms.ppt	Course	ceng 302	-
/~akilic	Lecturer	akilic	-
/~alan/METU-ISTEC/publications.htm	Lecturer	alan	-
/~alan/METU-ISTEC/publications/Bayir-Toroslu-Cosar.pdf	Lecturer	alan	-
/~alpaslan	Lecturer	alpaslan	-
/~alpaslan/fna.xml	Lecturer	alpaslan	-
/~alpaslan/main.swf	Lecturer	alpaslan	-
/~alpaslan/students.html	Lecturer	alpaslan	-
/~alpaslan/teaching.html	Lecturer	alpaslan	-
/~aykut	Lecturer	aykut	-
/~aykut/cv.pdf	Lecturer	aykut	-
/~aysegul	Lecturer	aysegul	-
/~aysun	Lecturer	aysun	-
/~birturk/birturk.html	Lecturer	birturk	-
/~birturk/birturk.html	Lecturer	birturk	-
/~bozsahin	Lecturer	bozsahin	-
/~bozsahin/abhofl.html	Lecturer	bozsahin	-
/~bozsahin/caltm.pdf	Lecturer	bozsahin	-
/~bozsahin/carg.pdf	Lecturer	bozsahin	-
/~bozsahin/dbbb.pdf	Lecturer	bozsahin	-
/~bozsahin/misc.html	Lecturer	bozsahin	-
/~bozsahin/nli/ceng563/lect/notes4a.pdf	Course	ceng 563	-
/~bozsahin/nli/ceng563/lect/notes4e.pdf	Course	ceng 563	-
/~bozsahin/nli/ceng584/ann/index.html	Course	ceng 584	-
/~bozsahin/nli/ceng584/lect/index.html	Course	ceng 584	-

Table A.2: (continued)

/~bozsahin/nli/ceng584/link/index.html	Course	ceng 584	-
/~bozsahin/research.html	Lecturer	bozsahin	-
/~bozsahin/schonfinkel.pdf	Lecturer	bozsahin	-
/~bozsahin/tpd-bci2003.pdf	Lecturer	bozsahin	-
/~bozsahin/wowofl.pdf	Lecturer	bozsahin	-
/~bozyigit	Lecturer	bozyigit	-
/~bozyigit/Courses/cng351/Lectures/Lec02_ secondaryStorageDevices.ppt	Course	ceng 351	-
/~bozyigit/CurrentCourses.html	Lecturer	bozyigit	-
/~cagatay	Lecturer	cagatay	-
/~ceng111/lab/grades/section5_ grades.html	Course	ceng 111	-
/~ceng111/lab/questions/q7.py	Course	ceng 111	-
/~ceng111/the4.pdf	Course	ceng 111	-
/~ceng140/the3.pdf	Course	ceng 140	-
/~cosar	Lecturer	cosar	-
/~cosar/556/Syllabus-556.pdf	Course	ceng 556	-
/~cuneyt	Lecturer	cuneyt	-
/~cuneyt/c_ cpp_ questions/c_ cpp_ questions_ tr.html	Lecturer	cuneyt	-
/~deniz	Lecturer	deniz	-
/~dogru	Lecturer	dogru	-
/~dogru/cose.pdf	Lecturer	dogru	-
/~dogru/oo9.pdf	Lecturer	dogru	-
/~dogru/ReqTemplate.doc	Lecturer	dogru	-
/~dogru/resume.html	Lecturer	dogru	-
/~dogru/se1.pdf	Lecturer	dogru	-
/~dogru/se2.pdf	Lecturer	dogru	-
/~dogru/se4.pdf	Lecturer	dogru	-
/~dogru/se5.pdf	Lecturer	dogru	-
/~dogru/se5.pdf	Lecturer	dogru	-



Table A.2: (continued)

/~e114068/guestbook6/gb.php	Student Page	-	-
/~e116471/designofcrycopro.pdf	Student Page	-	-
/~e1195288/CodeNotes_for_J2EE.pdf	Student Page	-	-
/~e1195288/Java%20Programming%20Un- leashed.pdf	Student Page	-	-
/~e120329	Student Page	-	-
/~e120329/counter.php	Student Page	-	-
/~e120329/pqstream_dkucuk.pdf	Student Page	-	-
/~e120329/wordle5.bmp	Student Page	-	-
/~e120346	Student Page	-	-
/~e120353/fong.pdf	Student Page	-	-
/~e120353/HoughTransform/FP_analysis.pdf	Student Page	-	-
/~e120353/HoughTransform/FP_analysis.pdf	Student Page	-	-
/~e1250133/kariyer.html	Student Page	-	-
/~e125043/gazeller.htm	Student Page	Music	-
/~e1250984/resume.pdf	Student Page	-	-

Table A.2: (continued)

/_e1272087	Student Page	-	-
/_e1297431/index.php?c=uludagsozluk&s=e/183/% 20% 20/contact.php	Student Page	-	-
/_e1297431/index.php?c=uludagsozluk&s=e/183/contact.php	Student Page	-	-
/_e1297431/index.php?c=uludagsozluk&s=e/contact.php	Student Page	-	-
/_e1321751/istenen/Trees.ppt	Student Page	-	-
/_e1347434	Student Page	-	-
/_e1347657	Student Page	-	-
/_e1389568/ozan_cv_april2008.pdf	Student Page	-	-
/_e1394618	Student Page	-	-
/_e1402668	Student Page	-	-
/_e1416056/hw4/guestbook/guestbook.php	Student Page	-	-
/_e1416056/hw4/guestbook/guestbook.php	Student Page	-	-
/_e1448380	Student Page	-	-
/_e1448596/seftali/update.rdf	Student Page	-	-
/_e1448786	Student Page	-	-

Table A.2: (continued)

/_e1448786/Evanescence-06-Understanding.MP3	Student Page	Music	-
/_e1448786/Evanescence% 20-% 2006% 20-% 20Understanding.MP3	Student Page	Music	-
/_e1448786/Evanescence% 20-% 2006% 20-% 20Understanding.MP3	Student Page	Music	-
/_e1448927/summer% 20practice% 20re-ports/ALARKO.doc	Student Page	Intern	-
/_e1449016/IEEE% 20830-1998% 20Recommended% 20Practice% 20for% 20Software% 20Requirements% 20Specifications.pdf	Student Page	-	-
/_e1449115/omertari	Student Page	-	-
/_e1449289	Student Page	-	-
/_e1449289/lig	Student Page	-	-
/_e1474022	Student Page	-	-
/_e1474022/android	Student Page	-	-
/_e1474022/android/Android% 20development% 20books	Student Page	Android	-
/_e1474022/android/Hello_ Android-Introducing_ Googles-Mobile_ Development_ Platform.pdf	Student Page	Android	-
/_e1502038	Student Page	-	-
/_e1502780	Student Page	-	-

Table A.2: (continued)

/~e1526581/Courses/CENG336/Lecture% 20Notes/ch2-2.ppt	Student Page	ceng 336	-
/~e1560044/melodiler	Student Page	Music	-
/~e1560176/files/algo/in	Student Page	Algororithm	-
/~e1560176/files/algo/out	Student Page	Algorithm	-
/~e1560200	Student Page	-	-
/~e1560200/PICos18_ tuto_ us.pdf	Student Page	-	-
/~e1560440/E_ Book/2- 1/Data/C++:HowTo....chm	Student Page	C++	-
/~e1595354/latex.pdf	Student Page	-	-
/~e1631191	Student Page	-	-
/~e1678879	Student Page	-	-
/~e1678879/Discrete.Mathematics.And.Its....pdf	Student Page	Discrete Math	-
/~erdas	Lecturer	erdas	-
/~erkut	Lecturer	erkut	-
/~erkut	Lecturer	erkut	-
/~erkut/cv.pdf	Lecturer	erkut	-
/~erkut/etv09.pdf	Lecturer	erkut	-
/~erman/java	Lecturer	erman	Java
/~erman/java/25eylul/25eylul.html	Lecturer	erman	Java
/~erman/java/sag.html	Lecturer	erman	Java

Table A.2: (continued)

/~erman/java/sol.html	Lecturer	erman	Java
/~erman/java/ust.html	Lecturer	erman	Java
/~erol	Lecturer	erol	-
/~ftitrek	Lecturer	ftitrek	-
/~futbol/2008/index.php?action=duy	Misc	Football	-
/~genc	Lecturer	genc	-
/~genc	Lecturer	genc	-
/~genc/334/Ch_ 1_ HC.ppt	Course	ceng 334	genc
/~genc/334/Ch_ 11_ Vista.ppt	Course	ceng 334	genc
/~genc/334/Ch_ 82_ MPS.ppt	Course	ceng 334	genc
/~genc/476/476.html	Course	ceng 476	genc
/~genc/index_ files/Page329.html	Lecturer	genc	-
/~genc/index_ files/Page372.html	Lecturer	genc	-
/~gokcen/tez.pdf	Lecturer	gokcen	-
/~gokdeniz	Lecturer	gokdeniz	-
/~gtumuklu	Lecturer	gtumuklu	-
/~gtumuklu/web/SE548/Reading% 20Material	Course	se 548	gtumuklu
/~gulen	Lecturer	gulen	-
/~isler	Lecturer	isler	-
/~isler/ceng732_ ComputerAnimation	Course	ceng 732	isler
/~ismet/cookbook	Lecturer	ismet	-
/~karagoz	Lecturer	karagoz	-
/~karagoz	Lecturer	karagoz	-
/~karagoz/ceng302/302-B+tree-ind-hash.pdf	Course	ceng 302	karagoz
/~karagoz/ceng302/302-B+tree-ind-hash.pdf	Course	ceng 302	karagoz
/~karagoz/ceng302/basic.pdf	Course	ceng 302	karagoz
/~karagoz/ceng302/FurtherDep.ppt	Course	ceng 302	karagoz
/~karagoz/ceng714-spr0809.htm	Course	ceng 714	karagoz
/~karagoz/ceng714/ceng714-fall05/paper-privacy-pres-mining.pdf	Course	ceng 714	karagoz

Table A.2: (continued)

/~karagoz/ceng770-fall2010-syllabus.pdf	Course	ceng 714	karagoz
/~karagoz/ceng770-fall2010-syllabus.pdf	Course	ceng 714	karagoz
/~karagoz/conf.html	Lecturer	karagoz	-
/~karagoz/interest.html	Lecturer	karagoz	-
/~kasim/reliability.pdf	Lecturer	kasim	-
/~ketenci	Lecturer	ketenci	-
/~ketenci/UserForm.php	Lecturer	ketenci	-
/~levent	Lecturer	levent	-
/~nafiz	Lecturer	nafiz	-
/~nafiz/papers/MStthesis.pdf	Lecturer	nafiz	-
/~nebil/index.html	Lecturer	nebil	-
/~nihan	Lecturer	nihan	-
/~nihan/ceng302	Course	ceng 302	nihan
/~nihan/ceng302/btrees.ppt	Course	ceng 302	nihan
/~nihan/ceng302/dbms.ppt	Course	ceng 302	nihan
/~nihan/ceng302/dbms.ppt	Course	ceng 302	nihan
/~nihan/ceng302/ER.ppt	Course	ceng 302	nihan
/~nihan/ceng302/index.htm	Course	ceng 302	nihan
/~nihan/ceng302/secondaryStorageDevices.ppt	Course	ceng 302	nihan
/~nihan/ceng302/secondaryStorageDevices.ppt	Course	ceng 302	nihan
/~nihan/ceng302/sequentialfiles.ppt	Course	ceng 302	nihan
/~nihan/ceng302/sequentialfiles.ppt	Course	ceng 302	nihan
/~nihan/CV2007.htm	Lecturer	nihan	-
/~nihan/Pub_list.htm	Lecturer	nihan	-
/~oguztuzn	Lecturer	oguztuzn	-
/~oguztuzn/courses	Lecturer	oguztuzn	-
/~oguztuzn/publications	Lecturer	oguztuzn	-
/~onur	Lecturer	onur	-
/~onur	Lecturer	onur	-
/~onur/diger.html	Lecturer	onur	-

Table A.2: (continued)

/~onur/english/index.html	Lecturer	onur	-
/~onur/site.pdf	Lecturer	onur	-
/~orald	Lecturer	orald	-
/~otopcu	Lecturer	otopcu	-
/~polat	Lecturer	polat	-
/~polat/educat.htm	Lecturer	polat	-
/~ruken	Lecturer	ruken	-
/~ruken	Lecturer	ruken	-
/~sciftci	Lecturer	sciftci	-
/~selma	Lecturer	selma	-
/~sener	Lecturer	sener	-
/~sercan	Lecturer	sercan	-
/~sertan/navigasyon72dpi.pdf	Lecturer	sertan	-
/~sertan/navigasyon72dpi.pdf	Lecturer	sertan	-
/~sibel	Lecturer	sibel	-
/~sibel	Lecturer	sibel	-
/~sibel/es303/mete_303_outline.html	Course	ceng 303	sibel
/~sibel/es303/Schema_303_new.pdf	Course	ceng 303	sibel
/~sibel/es303/st303index.html	Course	ceng 303	sibel
/~sibel/index.html	Lecturer	sibel	-
/~sibel/Listatiflar_tr4.pdf	Lecturer	sibel	-
/~sibel/papers/aslanthesis.pdf	Lecturer	sibel	-
/~sibel/stdevin.html	Lecturer	sibel	-
/~tcan	Lecturer	tcan	-
/~tcan	Lecturer	tcan	-
/~tcan/bin504_20101/Schedule/bin504_week11.pdf	Lecturer	tcan	-
/~tcan/bin504_20101/Schedule/bin504_week8.pdf	Lecturer	tcan	-
/~tcan/ceng465_s0809/overview.shtml	Course	ceng 465	tcan

Table A.2: (continued)

/_tcan/ceng465/Assignments/assignment2.pdf	Course	ceng 465	tcan
/_tcan/ceng465/Schedule/marray-intro3.pdf	Course	ceng 465	tcan
/_tcan/ceng732/Schedule/ceng732_ week2.pdf	Course	ceng 732	tcan
/_tcan/ceng734_ 20101	Course	ceng 734	tcan
/_tcan/ceng734_ 20101/Schedule/FunctionallyGuidedAlignment	Course	ceng 734	tcan
/_tcan/ceng734_ 20101/Schedule/index.shtml	Course	ceng 734	tcan
/_tcan/ceng734/Schedule/week1.pdf	Course	ceng 734	tcan
/_tcan/fpv	Lecturer	tcan	-
/_tcan/ProteinNetworkPapers.html	Lecturer	tcan	-
/_tcan/publications.html	Lecturer	tcan	-
/_tcan/publications.html	Lecturer	tcan	-
/_tcan/publications/BIO-121.pdf	Lecturer	tcan	-
/_tcan/se705_ s0809/Schedule/se705_ week13.pdf	Course	ceng 705	tcan
/_tcan/se705_ s0910/Schedule/SamplePhase2_ 2.pdf	Course	ceng 705	tcan
/_tcan/se705_ s0910/Schedule/se705_ week3.pdf	Course	ceng 705	tcan
/_tcan/se705/Schedule/assignment6.pdf	Course	ceng 705	tcan
/_tcan/se705/Schedule/week12_ speech.pdf	Course	ceng 705	tcan
/_tcan/tolgacan-cv.pdf	Lecturer	tcan	-
/_toroslu	Lecturer	toroslu	-
/_ucoluk/bm	Lecturer	ucoluk	-
/_ucoluk/bm.html	Lecturer	ucoluk	-
/_ucoluk/darwin/node10.html	Lecturer	ucoluk	-
/_ucoluk/darwin/node4.html	Lecturer	ucoluk	-
/_ucoluk/research/lisp/lispman/lispman.html	Lecturer	ucoluk	-
/_ucoluk/research/publications/tsp.pdf	Lecturer	ucoluk	-
/_ucoluk/research/publications/tsp.pdf	Lecturer	ucoluk	-
/_ucoluk/research/publications/tspnew.pdf	Lecturer	ucoluk	-



Table A.2: (continued)

/~ucoluk/say.cgi	Lecturer	ucoluk	-
/~ucoluk/say.php	Lecturer	ucoluk	-
/~ucoluk/yazin/arif_ erkan.html	Lecturer	ucoluk	-
/~ucoluk/yazin/bilim3.html	Lecturer	ucoluk	-
/~ucoluk/yazin/marr.html	Lecturer	ucoluk	-
/~ucoluk/yazin/OP_ ATK.html	Lecturer	ucoluk	-
/~vbi	Misc	-	-
/~vbi	Misc	-	-
/~volkan	Lecturer	volkan	-
/~volkan	Lecturer	volkan	-
/~volkan/VAtalay-Publications.html	Lecturer	volkan	-
/~yalabik	Lecturer	yalabik	-
/~yazici	Lecturer	yazici	-
/~yazici/header.htm	Lecturer	yazici	-
/~yazici/menu.htm	Lecturer	yazici	-
/about/about	Misc	-	-
/about/about	Misc	-	-
/about/contact	Misc	-	-
/about/location	Misc	-	-
/box/sitemap	Misc	-	-
/contact.php	Misc	-	-
/contest/upem	Misc	-	-
/course/ceng111/faculty	Course	ceng 111	-
/course/ceng111/lab	Course	ceng 111	-
/course/ceng111/library	Course	ceng 111	-
/Courses/?course=ceng300	Course	ceng 300	-
/Courses/?semester=20092	Course	All	-
/Courses/?semester=20092& course=ceng336& cedit=0	Course	ceng 336	-

Table A.2: (continued)

/Courses/?semester=20092& course=ceng382& cedit=0	Course	ceng 382	-
/Courses/?semester=20092& course=ceng567& cedit=0	Course	ceng 567	-
/courses/ceng232/2008/exp5.pdf	Course	ceng 232	-
/courses/ceng242	Course	ceng 242	-
/courses/ceng242/assignments	Course	ceng 242	-
/courses/ceng242/documents/sebesta/Ch3part2.pdf	Course	ceng 242	-
/courses/ceng242/documents/slides/binding.pdf	Course	ceng 242	-
/courses/ceng242/main.html	Course	ceng 242	-
/courses/ceng242/menu.html	Course	ceng 242	-
/courses/ceng280	Course	ceng 280	-
/courses/ceng280/csMain.html	Course	ceng 280	-
/courses/ceng280/csToolbar.html	Course	ceng 280	-
/courses/ceng334/Ch_ 10_ UNIX.ppt	Course	ceng 334	-
/courses/ceng334/Ch_ 23_ IPC.ppt	Course	ceng 334	-
/courses/ceng334/Ch_ 24_ Deadlocks.ppt	Course	ceng 334	-
/courses/ceng334/Ch_ 5_ IO.ppt	Course	ceng 334	-
/courses/ceng336/2005/_ documents/timers.pdf	Course	ceng 336	-
/courses/ceng336/2005/_ documents/timers.pdf	Course	ceng 336	-
/courses/ceng351/assignments/index.html	Course	ceng 351	-
/courses/ceng351/documents/week1_ Introduc- tion_ section2.pdf	Course	ceng 351	-
/courses/ceng351/documents/week3_ Sequen- tialFiles_ section2.pdf	Course	ceng 351	-
/courses/ceng351/documents/week3_ Sequen- tialFiles_ section2.pdf	Course	ceng 351	-
/courses/ceng352/assignments/hw2.pdf	Course	ceng 352	-
/courses/ceng444	Course	ceng 444	-

Table A.2: (continued)

/courses/ceng444/csMain.html	Course	ceng 444	-
/courses/ceng444/csToolbar.html	Course	ceng 444	-
/courses/ceng444/lect/notes3b.pdf	Course	ceng 444	-
/courses/ceng444/lect/notes8.pdf	Course	ceng 444	-
/courses/ceng444/link/444_ phase2_ recitation.pdf	Course	ceng 444	-
/courses/ceng463	Course	ceng 463	-
/courses/ceng466	Course	ceng 466	-
/courses/ceng466/2007/frames/main.htm	Course	ceng 466	-
/courses/ceng466/2007/frames/syllabus.htm	Course	ceng 466	-
/courses/ceng469	Course	ceng 469	-
/courses/ceng469/2008/2006/documents / Suggested ProjectTopics.pdf	Course	ceng 469	-
/courses/ceng469/frames/main.htm	Course	ceng 469	-
/courses/ceng469/MenuFrame.htm	Course	ceng 469	-
/courses/ceng477	Course	ceng 477	-
/courses/ceng477	Course	ceng 477	-
/courses/ceng477/2004/documents/Illumination% 20Models% 20and% 20Surface% 20Rendering% 20Methods.pdf	Course	ceng 477	-
/courses/ceng477/2005/documents/uwashington-ray-tracing.pdf	Course	ceng 477	-
/courses/ceng477/2006/documents/lecturenotes_2006/week13_VisibleSurfaceDetection.ppt	Course	ceng 477	-
/courses/ceng477/2008/documents/lecturenotes_2008/week2.pdf	Course	ceng 477	-
/courses/ceng477/assignments/index.html	Course	ceng 477	-
/courses/ceng477/documents	Course	ceng 477	-
/courses/ceng477/documents/index.html	Course	ceng 477	-
/courses/ceng477/documents/index.html	Course	ceng 477	-

Table A.2: (continued)

/courses/ceng477/documents/lnotes/week1.pdf	Course	ceng 477	-
/courses/ceng477/documents/lnotes/week1.ppt	Course	ceng 477	-
/courses/ceng477/documents/lnotes/week6.ppt	Course	ceng 477	-
/courses/ceng477/documents/lnotes/week6.ppt	Course	ceng 477	-
/courses/ceng477/links.html	Course	ceng 477	-
/courses/ceng477/main.html	Course	ceng 477	-
/courses/ceng477/main.html	Course	ceng 477	-
/courses/ceng477/menu.html	Course	ceng 477	-
/courses/ceng477/menu.html	Course	ceng 477	-
/courses/ceng490	Course	ceng 490	-
/courses/ceng490	Course	ceng 490	-
/courses/ceng490/documents	Course	ceng 490	-
/courses/ceng490/documents/491syllabus Fall2010.html	Course	ceng 490	-
/courses/ceng490/documents/CEng491- ProjectGroups_files/sheet001.html	Course	ceng 490	-
/courses/ceng490/documents/CEng491- ProjectGroups_files/tabstrip.html	Course	ceng 490	-
/courses/ceng490/documents/CEng491- ProjectGroups.html	Course	ceng 490	-
/courses/ceng490/documents/CEng491- ProjectGroups.html	Course	ceng 490	-
/courses/ceng490/documents/Project_Presenta- tion_Minder_Yazilim_BCI-InAction.ppt	Course	ceng 490	-
/courses/ceng490/documents/syllabus-492.html	Course	ceng 490	-
/courses/ceng490/finaldemo.html	Course	ceng 490	-
/courses/ceng490/main.html	Course	ceng 490	-
/courses/ceng490/main.html	Course	ceng 490	-
/courses/ceng490/menu.html	Course	ceng 490	-
/courses/ceng490/menu.html	Course	ceng 490	-

Table A.2: (continued)

/courses/ceng536	Course	ceng 536	-
/courses/ceng536/documents/sp_ signals.pdf	Course	ceng 536	-
/courses/ceng536/homeworks	Course	ceng 536	-
/courses/ceng536/main.html	Course	ceng 536	-
/courses/ceng536/menu.html	Course	ceng 536	-
/courses/ceng536/sources	Course	ceng 536	-
/courses/ceng536/sources/uml/uml.pdf	Course	ceng 536	-
/courses/ceng564	Course	ceng 564	-
/courses/ceng564/assg.html	Course	ceng 564	-
/courses/ceng564/contents.htm	Course	ceng 564	-
/courses/ceng564/main.html	Course	ceng 564	-
/courses/ceng574	Course	ceng 574	-
/courses/ceng574/CENG574-syllabus-Fall10.html	Course	ceng 574	-
/courses/ceng701	Course	ceng 701	-
/courses/ceng713	Course	ceng 713	-
/courses/ceng713/assignments	Course	ceng 713	-
/courses/ceng713/documents/parallelea.pdf	Course	ceng 713	-
/courses/ceng713/main.html	Course	ceng 713	-
/courses/ceng713/menu.html	Course	ceng 713	-
/Courses/homework.php?hid=1287	Misc	-	-
/courses/secondprog/ceng707/CENG707 syl- labus.pdf	Course	ceng 707	-
/courseweb/ceng242	Course	ceng 242	-
/courseweb/ceng242/main.html	Course	ceng 242	-
/courseweb/ceng242/syllabus.html	Course	ceng 242	-
/doc/index?idx=doc:index	Document	-	-
/doc/services/certs	Document	-	-
/doc/services/certs/certxp	Document	-	-
/doc/services/email/thunderbird	Document	-	-

Table A.2: (continued)

/doc/services/index	Document	-	-
/doc/services/index	Document	-	-
/doc/services/news/thunderbird	Document	-	-
/doc/services/quota	Document	-	-
/doc/services/sieve	Document	-	-
/doc/services/web	Document	-	-
/doc/studentdoc/index	Document	-	-
/feed.php	Feedback	-	-
/feed.php	Feedback	-	-
/feed.php?mode=list& ns=	Feedback	-	-
/feed.php?mode=list& ns=tanitim	Feedback	-	-
/grad/courses	Grad	-	-
/grad/curriculum	Grad	-	-
/grad/index	Grad	-	-
/grad/mswotceng	Grad	-	-
/grad/mswotceng	Grad	-	-
/grad/mswotse	Grad	-	-
/grad/phdqual	Grad	-	-
/hw4/guestbook/h% 3C/% 3C/td	Misc	-	-
/hw4/guestbook/h% 3C/td	Misc	-	-
/index	Misc	-	-
/index?do=index	Misc	-	-
/index.php?id=news/20101/courses/ sched- ule.html& purge=1	Misc	News	-
/index.php?id=news/seminar& semtab=2005	Seminar	News	-
/index.php?id=news/seminar& semtab=2006	Seminar	News	-
/index.php?id=news/seminar& semtab=subscribe	Seminar	News	-
/index.php?id=news/seminar& semtab=Upcoming	Seminar	News	-

Table A.2: (continued)

/index.php?id=undergrad/courses&crsyear=20091	Undergrad	News	-
/index.php?option=com_content&task=section& id=1& Itemid=105	Misc	News	-
/index.php?option=com_content& task=view& id=185& Itemid=105	Misc	News	-
/index.php?option=com_content& task=view& id=34& Itemid=77	Misc	News	-
/index.php?option=com_content& task=view& id=34& Itemid=77	Misc	News	-
/index.php?option=com_content& task=view& id=36& Itemid=54	Misc	News	-
/index.php?option=com_content& task=view& id=45& Itemid=62	Misc	News	-
/index.php?option=com_content& task=view& id=54& Itemid=105	Misc	News	-
/index.php?option=com_cow_courses&task=view& semester=& course=ceng230	Course	ceng 230	News
/index.php?option=com_cow_courses&task=view& semester=& course=ceng520	Course	ceng 520	News
/index.php?option=com_cow_docs& category=0& Itemid=102	Misc	News	-
/index.php?option=com_cow_people&task=list& type=staff& group=0& Itemid=70	Misc	News	-
/index.php?option=com_cow_people&task=list& type=staff& group=0& Itemid=70	Misc	News	-
/index.php?option=com_cow_people&task=view& type=staff& group=0& user-name=birturk	Lecturer	birturk	-

Table A.2: (continued)

/index.php?option=com_cow_people&task=view&type=staff&group=0&username=cosar	Lecturer	cosar	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=dogru	Lecturer	dogru	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=erol	Lecturer	erol	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=karagoz	Lecturer	karagoz	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=karagoz	Lecturer	karagoz	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=oguztuzn	Lecturer	oguztuzn	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=ruken	Lecturer	ruken	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=sener	Lecturer	sener	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=skalkan	Lecturer	skalkan	-
/index.php?option=com_cow_people&task=view&type=staff&group=0&username=volkan	Lecturer	volkan	-



Table A.2: (continued)

/index.php?option=com_cow_people&task=view&type=staff&group=0&username=yazici	Lecturer	yazici	-
/index.php?option=com_cow_seminars&limitstart=0&task=view	Misc	News	-
/index.php?option=com_cow_seminars&type=all	Seminar	-	-
/index.php?printview=1	Misc	-	-
/index.php?printview=1	Misc	-	-
/index.php?purge=1	Misc	-	-
/index.tr?rev=1295437556&do=diff	Misc	-	-
/index2.php?option=com_rssxt&type=RSS&no_html=1&cat=Events	Misc	-	-
/index2.php?option=com_rssxt&type=RSS&no_html=1&cat=News	Misc	News	-
/index2.php?option=com_rssxt&type=RSS&no_html=1&cat=News	Misc	News	-
/indonesia.htm	Misc	-	-
/ineks.html	Misc	-	-
/lib/exe/ajax.php	Library	-	-
/lib/exe/css.php?s=all&t=arctic&tseed=1277920437	Library	-	-
/lib/exe/css.php?s=all&t=arctic&tseed=1278738978	Library	-	-
/lib/exe/css.php?s=all&t=arctic&tseed=1279298329	Library	-	-
/lib/exe/css.php?s=all&t=arctic&tseed=1279806924	Library	-	-
/lib/exe/css.php?s=all&t=arctic&tseed=1282774372	Library	-	-

Table A.2: (continued)

/lib/exe/css.php?s=all& tseed=1282774372	t=arctic&	Library	-	-
/lib/exe/css.php?s=all& tseed=1285257405	t=arctic&	Library	-	-
/lib/exe/css.php?s=all& tseed=1285257405	t=arctic&	Library	-	-
/lib/exe/css.php?s=all& tseed=1295439608	t=arctic&	Library	-	-
/lib/exe/css.php?s=all& tseed=1295439608	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1277920437	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1278738978	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1279298329	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1279806924	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1282774372	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1282774372	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1285257405	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1285257405	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1295439608	t=arctic&	Library	-	-
/lib/exe/css.php?s=print& tseed=1295439608	t=arctic&	Library	-	-

Table A.2: (continued)

/lib/exe/css.php?t=arctic& tseed=1277920437	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1278738978	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1279298329	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1279806924	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1282774372	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1282774372	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1285257405	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1285257405	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1295439608	Library	-	-
/lib/exe/css.php?t=arctic& tseed=1295439608	Library	-	-
/lib/exe/js.php?tseed=1277920437& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1278738978& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1279298329& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1279806924& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1282774372& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1282774372& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1285257405& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1285257405& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1295439608& amp;t=arctic	Library	-	-
/lib/exe/js.php?tseed=1295439608& t=arctic	Library	-	-
/lib/exe/js.php?tseed=1295439608& t=arctic	Library	-	-
/lib/exe/opensearch.php	Library	-	-
/lib/exe/opensearch.php	Library	-	-
/lib/plugins/cow/courses.ajax.php	Library	Plugin	-
/lib/plugins/cow/courses.ajax.php	Library	Plugin	-
/lib/plugins/cow/ical.php?dtstart=-1month& seminar& /seminar.ics	Library	Plugin	-
/lib/plugins/cow/ical.php?dtstart=-1year& /cow.ics	Library	Plugin	-

Table A.2: (continued)

/lib/plugins/cow/ical.php?dtstart=-1year& seminar& /seminar.ics	Library	Plugin	-
/lib/plugins/cow/ical.php?dtstart=-3months& hw& /hw.ics	Library	Plugin	-
/lib/plugins/cow/ical.php?recache=true& dtstart=-1year& exams& /exams.ics	Library	Plugin	-
/lib/plugins/cow/seminar.ajax.php	Library	Plugin	-
/lib/plugins/indexmenu/ajax.php	Library	Plugin	-
/lib/plugins/randompage/ajax.php	Library	Plugin	-
/metu?do=search	Misc	-	-
/news/20091/acm_ contest/acm_ contest	Misc	News	-
/news/20092/courses/schedule	Course	All	-
/news/20093/arastirmagorevlisi	Misc	News	-
/news/20093/arastirmagorevlisi	Misc	News	-
/news/20093/facultypositions	Misc	News	-
/news/20093/facultypositions	Misc	News	-
/news/20093/icub	Misc	News	-
/news/20093/upem	Misc	News	-
/news/20093/upem	Misc	News	-
/news/20101/courses/announcement	Misc	News	-
/news/20101/courses/schedule.html	Misc	News	-
/news/20101/degerlendirme	Misc	News	-
/news/20101/degerlendirme2	Misc	News	-
/news/20101/doktorayeterlik	Misc	News	-
/news/20101/evraklar	Misc	News	-
/news/20101/ondegerlendirme	Misc	News	-
/news/20101/qual	Misc	News	-
/news/index	Misc	News	-
/news/nntp?semtab=metu.ceng.announce.admin& semid=424	Seminar	News	-

Table A.2: (continued)

/news/nntp?semtab=metu.ceng.announce.admin&semid=435	Seminar	News	-
/news/nntp?semtab=metu.ceng.announce&semid=914	Seminar	News	-
/news/nntp?semtab=Recent	Seminar	News	-
/news/seminar	Seminar	News	-
/news/seminar?semid=345	Seminar	News	-
/news/seminar?semid=346	Seminar	News	-
/news/seminar?semtab=subscribe	Seminar	News	-
/news/seminar?semtab=Upcoming	Seminar	News	-
/News/thread.php?group=metu.ceng.announce.admin	Misc	News	-
/News/thread.php?group=metu.ceng.announce.admin	Misc	News	-
/News/thread.php?group=metu.ceng.announce.jobs	Misc	News	-
/News/thread.php?group=metu.ceng.announce.sales	Misc	News	-
/News/thread.php?group=metu.ceng.announce.sales	Misc	News	-
/News/thread.php?group=metu.ceng.course.140	Course	ceng 140	News
/News/thread.php?group=metu.ceng.course.140	Course	ceng 140	News
/News/thread.php?group=metu.ceng.course.232	Course	ceng 232	News
/News/thread.php?group=metu.ceng.course.242	Course	ceng 242	News
/News/thread.php?group=metu.ceng.course.280	Course	ceng 280	News
/News/thread.php?group=metu.ceng.course.334	Course	ceng 334	News
/News/thread.php?group=metu.ceng.course.334	Course	ceng 334	News
/News/thread.php?group=metu.ceng.course.336	Course	ceng 336	News
/News/thread.php?group=metu.ceng.course.336	Course	ceng 336	News
/News/thread.php?group=metu.ceng.course.350	Course	ceng 350	News

Table A.2: (continued)

/News/thread.php?group=metu.ceng.course.382	Course	ceng 382	News
/News/thread.php?group=metu.ceng.course.436	Course	ceng 436	News
/News/thread.php?group=metu.ceng.course.443	Course	ceng 443	News
/News/thread.php?group=metu.ceng.course.462	Course	ceng 462	News
/News/thread.php?group=metu.ceng.course.463	Course	ceng 463	News
/News/thread.php?group=metu.ceng.course.465	Course	ceng 465	News
/News/thread.php?group=metu.ceng.course.465	Course	ceng 465	News
/News/thread.php?group=metu.ceng.course.483	Course	ceng 483	News
/News/thread.php?group=metu.ceng.course.562	Course	ceng 562	News
/News/thread.php?group=metu.ceng.course.567	Course	ceng 567	News
/News/thread.php?group=metu.ceng.course.568	Course	ceng 568	News
/News/thread.php?group=metu.ceng.deer	Misc	News	-
/News/thread.php?group=metu.ceng.deer	Misc	News	-
/News/thread.php?group=metu.ceng.kult.dizi	Misc	News	-
/News/thread.php?group=metu.ceng.kult.muzik	Misc	Music	News
/News/thread.php?group=metu.ceng.news	Misc	News	-
/News/thread.php?group=metu.ceng.news	Misc	News	-
/News/thread.php?group=metu.ceng.others. bunalim	Misc	News	-
/News/thread.php?group=metu.ceng.others. hardware	Misc	News	-
/News/thread.php?group=metu.ceng.others.zen	Misc	News	-
/News/thread.php?group=metu.ceng. second- prog.567	Course	ceng 567	News
/News/thread.php?group=metu.ceng. second- prog.se705	Course	se 705	News
/News/thread.php?group=metu.ceng.ses	Misc	News	-
/News/thread.php?group=metu.ceng.sports	Misc	News	-
/News/thread.php?group=metu.ceng.student. freshman	Misc	News	-

Table A.2: (continued)

/News/thread.php?group=metu.ceng.student. junior	Misc	News	-
/News/thread.php?group=metu.ceng.student. senior	Misc	News	-
/News/thread.php?group=metu.ceng.student. senior	Misc	News	-
/News/thread.php?group=metu.ceng.student. sophomore	Misc	News	-
/News/thread.php?group=metu.ceng.student. sophomore	Misc	News	-
/News/thread.php?group=metu.ceng.test	Misc	News	-
/News/thread.php?group=metu.ceng.test	Misc	News	-
/News/thread.php?group=metu.ceng.turnuva. futbol	Misc	News	-
/News/thread.php?group=metu.ceng.unix	Misc	News	-
/people/alumni/aykut/index	Student Page	Alumni	-
/people/alumni/e1347657/index	Student Page	Alumni	-
/people/alumni/guide	Student Page	Alumni	-
/people/alumni/index	Student Page	Alumni	-
/people/alumni/index	Student Page	Alumni	-
/people/alumni/index?do=edit	Student Page	Alumni	-
/people/alumni/index?do=edit& rev=	Student Page	Alumni	-

Table A.2: (continued)

/people/alumni/index?do=revisions	Student Page	Alumni	-
/people/assistants/index	Misc	-	-
/people/assistants/index	Misc	-	-
/people/faculty/genc/index	Lecturer	genc	-
/people/faculty/index	Lecturer	faculty	-
/people/faculty/index	Lecturer	faculty	-
/people/faculty/karagoz/index	Lecturer	karagoz	-
/people/faculty/sener/index	Lecturer	sener	-
/people/faculty/skalkan/index	Lecturer	skalkan	-
/people/faculty/skalkan/index	Lecturer	skalkan	-
/people/faculty/ucoluk/index	Lecturer	vural	-
/people/faculty/volkan/index	Lecturer	vural	-
/people/faculty/vural/index	Lecturer	vural	-
/people/index	Misc	-	-
/people/index?idx=people	Misc	-	-
/people/staff/index	Misc	-	-
/research/bioinfo/index?bibentry=can6& bibid=bioinfo.bib	Research	-	-
/research/graphics/index	Research	-	-
/research/grid/index	Research	-	-
/research/index	Research	-	-
/research/index	Research	-	-
/research/index?idx=research	Research	-	-
/research/kovan/index	Research	-	-
/research/mining/index	Research	-	-
/research/mining/index	Research	-	-
/research/parallel/index	Research	-	-
/senior/index	Misc	-	-
/senior/index	Misc	-	-



Table A.2: (continued)

/start?idx=start	Misc	-	-
/Student/homeworks.php	Student Page	-	-
/Student/homeworks.php?	Student Page	-	-
/Student/homeworks.php?hid=1287	Student Page	-	-
/Student/stajBilgileri.php	Student Page	Intern	-
/Student/stajBilgileri.php?task_ student_ staj_ info=add	Student Page	Intern	-
/tanitim/2001sunum/index.tr	Introduction	-	-
/tanitim/2008sunum/index.tr	Introduction	-	-
/tanitim/bmhakkinda.tr	Introduction	-	-
/tanitim/index.tr	Introduction	-	-
/undergrad/courses	Undergrad	-	-
/undergrad/courses	Undergrad	-	-
/undergrad/courses?crsprogram=all	Undergrad	-	-
/undergrad/curriculum	Undergrad	-	-
/undergrad/curriculum.tr	Undergrad	-	-
/undergrad/index	Undergrad	-	-
/undergrad/index	Undergrad	-	-
/undergrad/index?idx=undergrad	Undergrad	-	-
/undergrad/index.tr	Undergrad	-	-
http://www.ceng.metu.edu.tr/%7Ee1416056/hw4/guestbook/guestbook.php	Misc	-	-
http://www.ceng.metu.edu.tr/%7Ee1416056/hw4/guestbook/guestbook.php	Misc	-	-

## APPENDIX B

### ADDITIONAL RESULTS and EXPERIMENTS

We run our experiments with 3-fold, 5-fold and finally 10-fold cross validation. As a result, we obtain for each validation the average of each iteration. In this chapter, detailed results of each run is given.

#### B.1 3-Fold Cross Validation Detailed Results

##### B.1.1 Ksim Similarity Measures

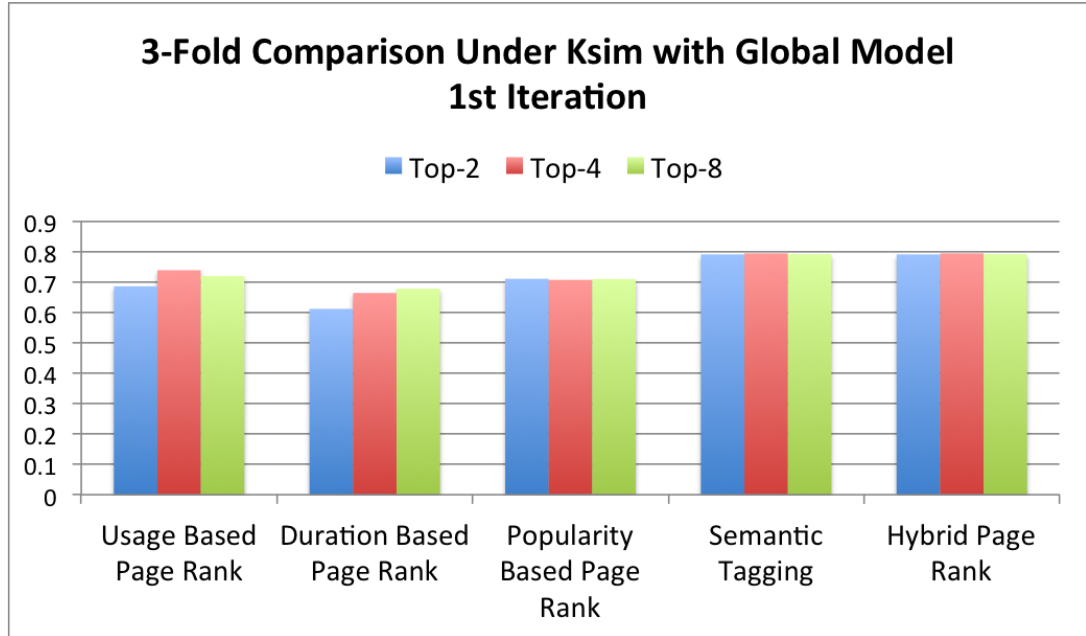


Figure B.1: First Iteration of 3-Fold Validation with Global Model Under Ksim Similarity

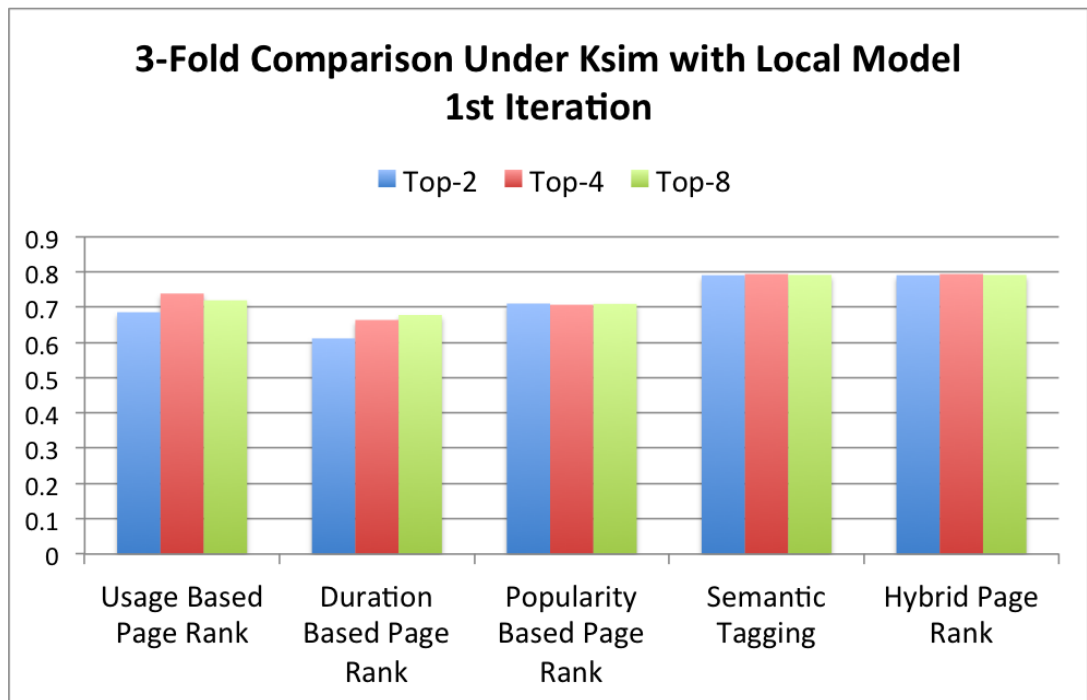


Figure B.2: First Iteration of 3-Fold Validation with Local Model Under Ksim Similarity

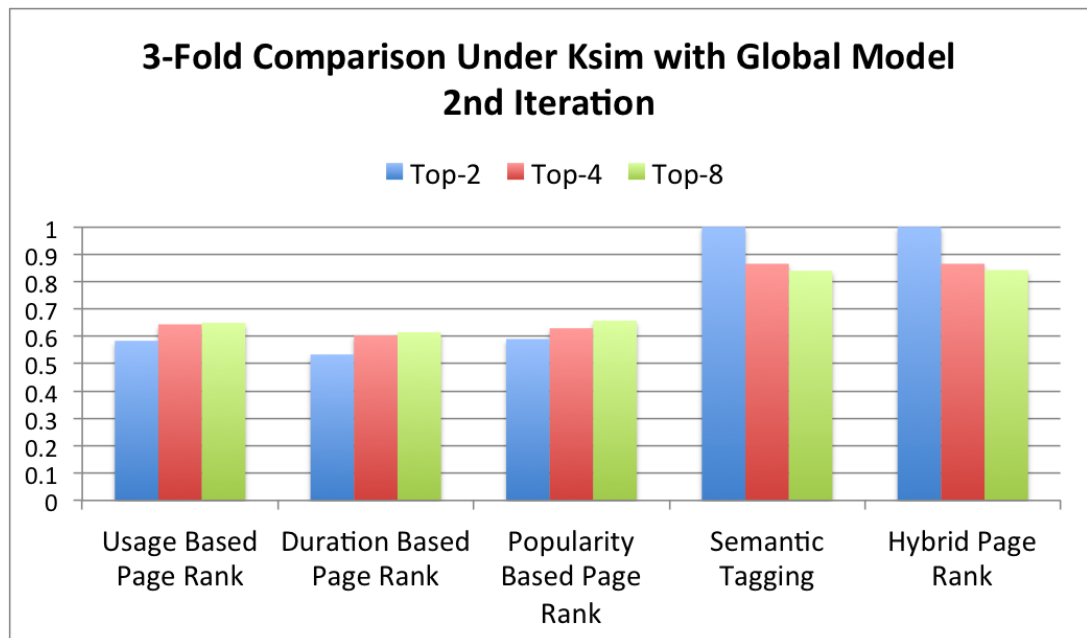


Figure B.3: Second Iteration of 3-Fold Validation with Global Model Under Ksim Similarity

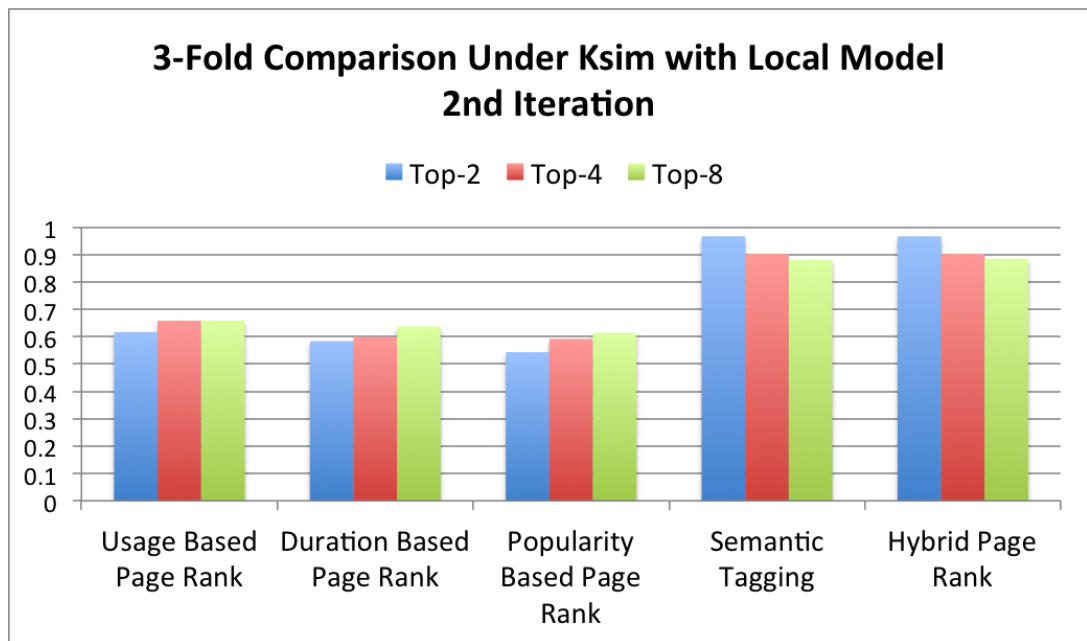


Figure B.4: Second Iteration of 3-Fold Validation with Local Model Under Ksim Similarity

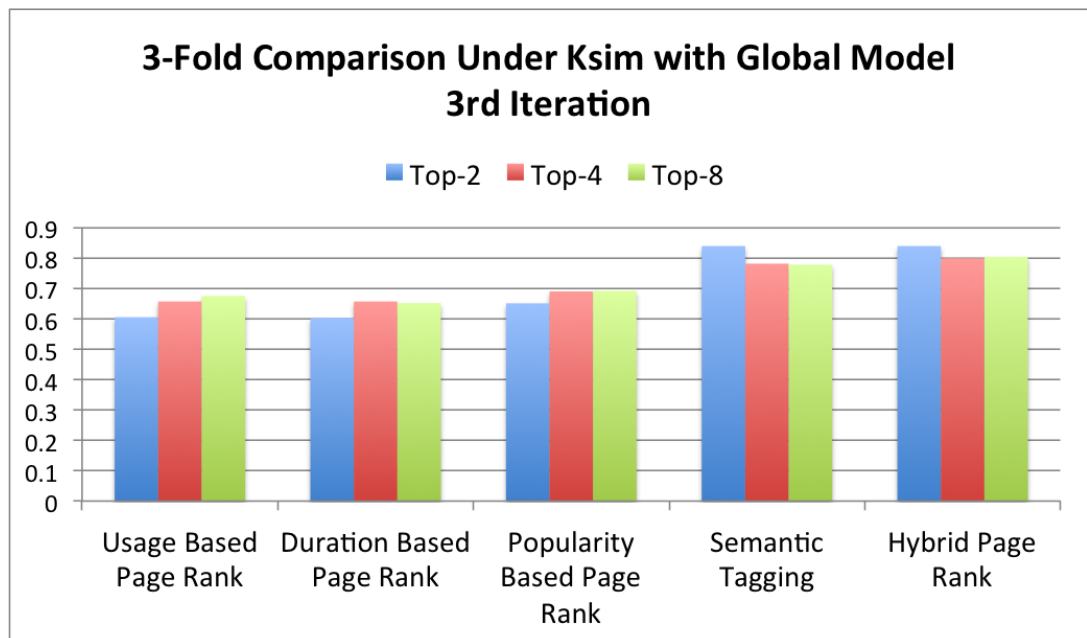


Figure B.5: Third Iteration of 3-Fold Validation with Global Model Under Ksim Similarity

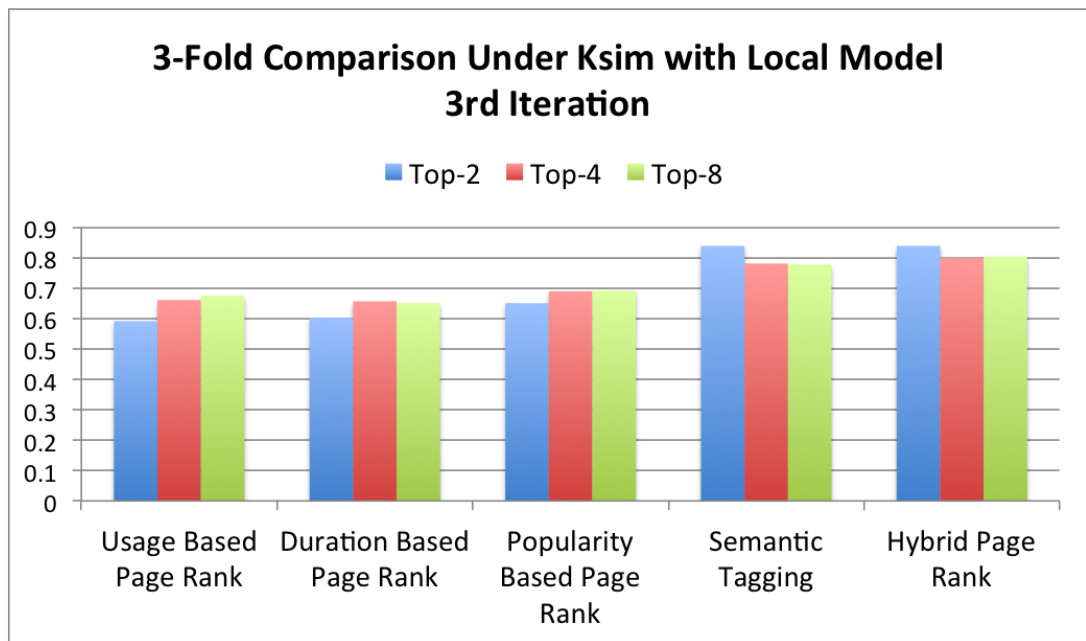


Figure B.6: Third Iteration of 3-Fold Validation with Local Model Under Ksim Similarity

**B.1.2 Osim Similarity Measures**

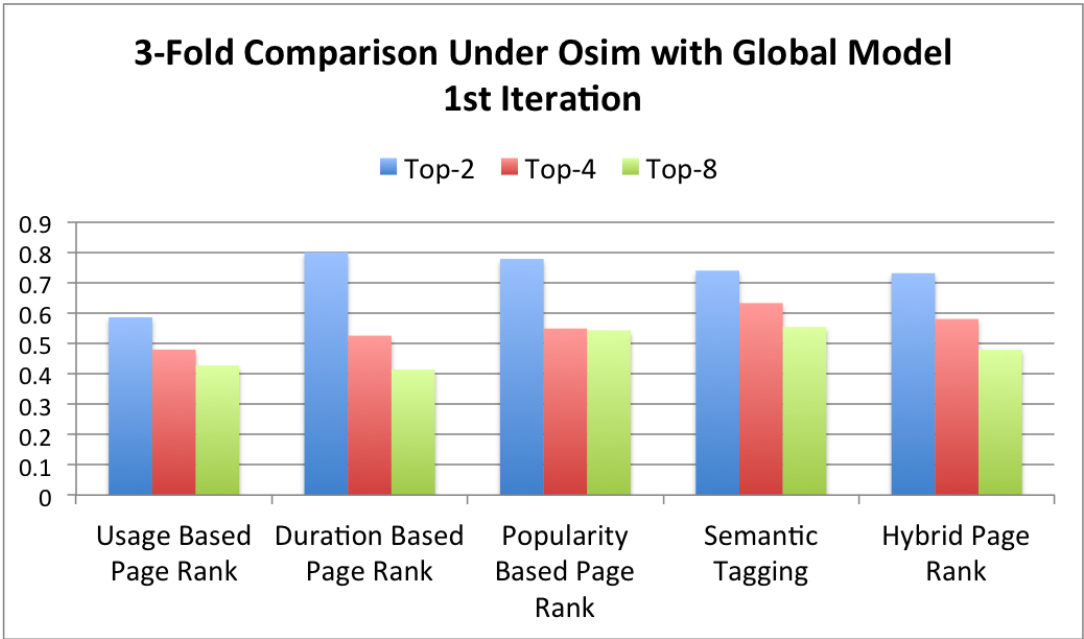


Figure B.7: First Iteration of 3-Fold Validation with Global Model Under Osim Similarity

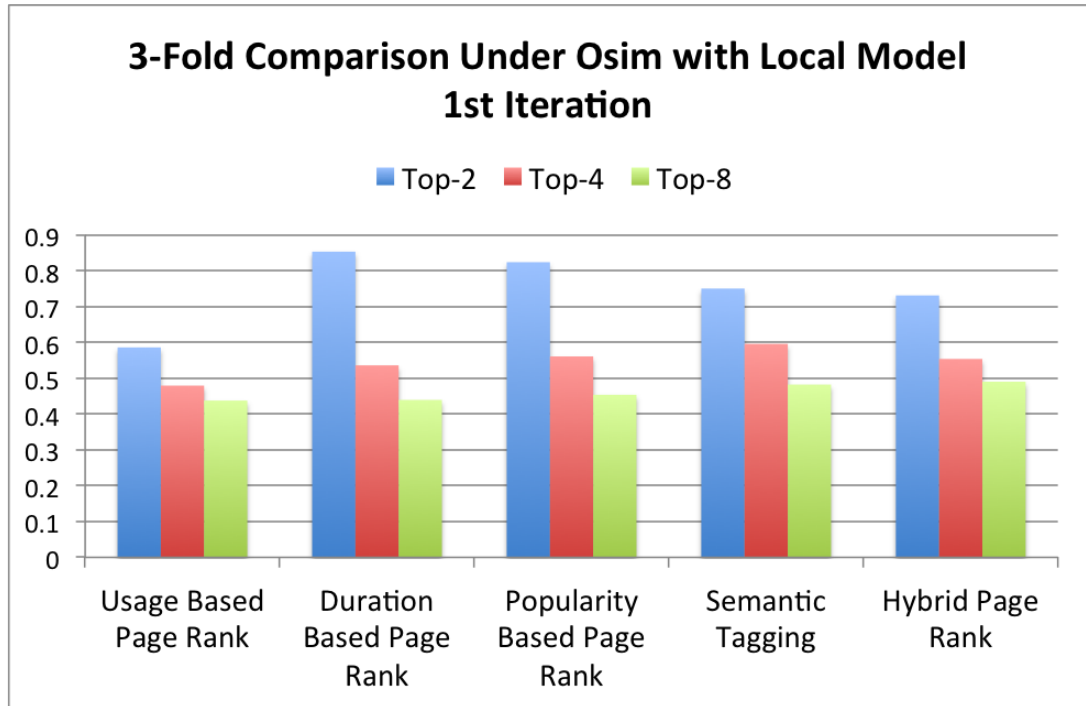


Figure B.8: First Iteration of 3-Fold Validation with Local Model Under Osim Similarity

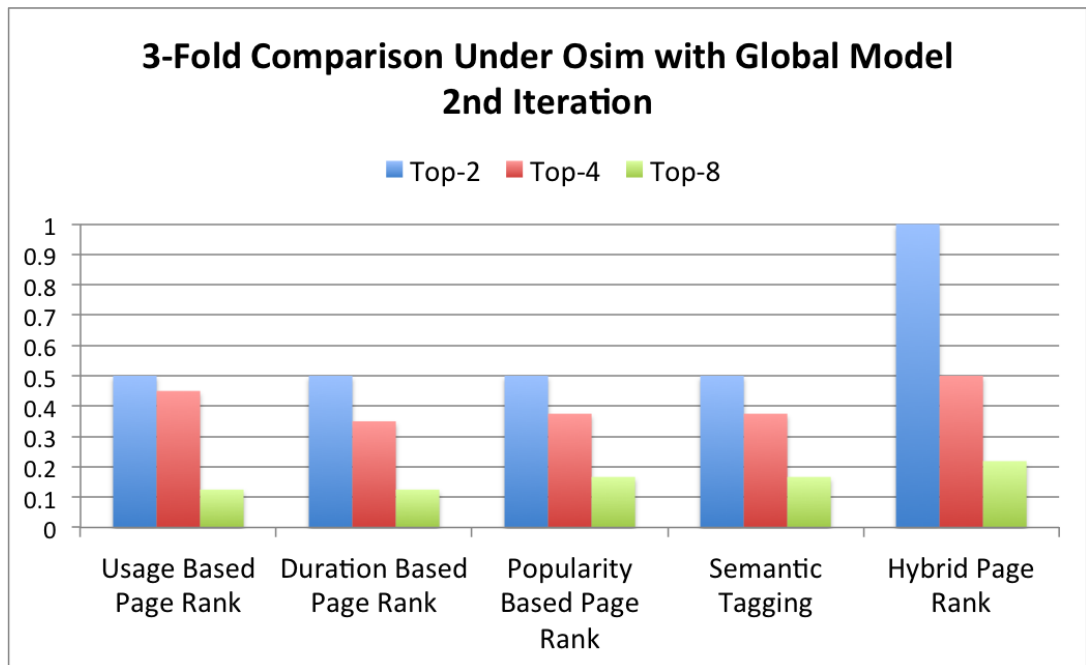


Figure B.9: Second Iteration of 3-Fold Validation with Global Model Under Osim Similarity

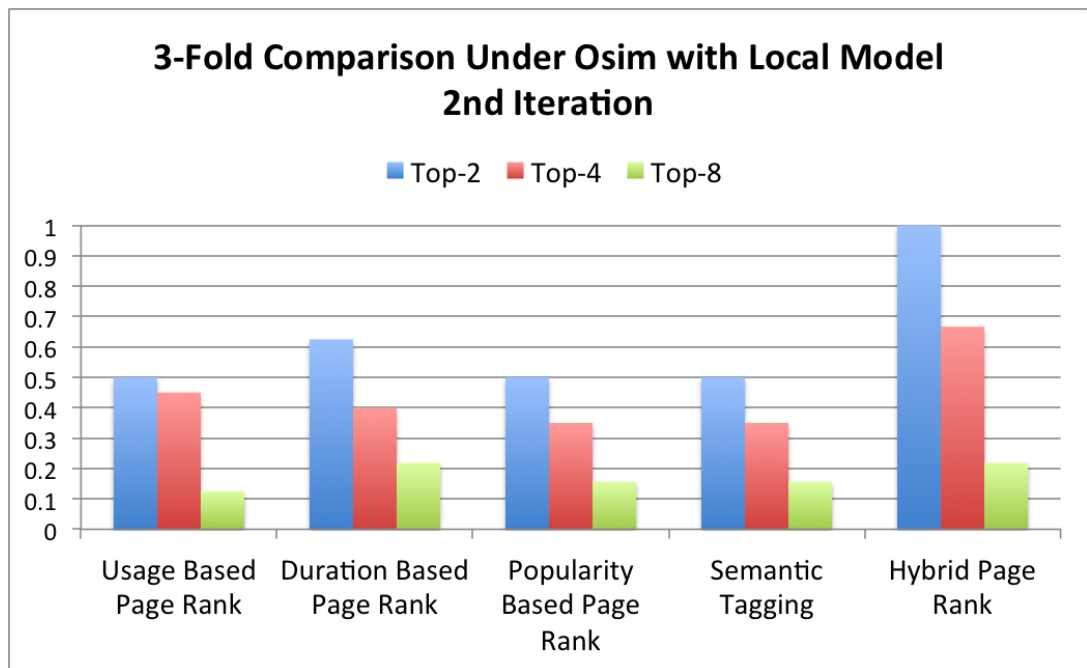


Figure B.10: Second Iteration of 3-Fold Validation with Local Model Under Osim Similarity

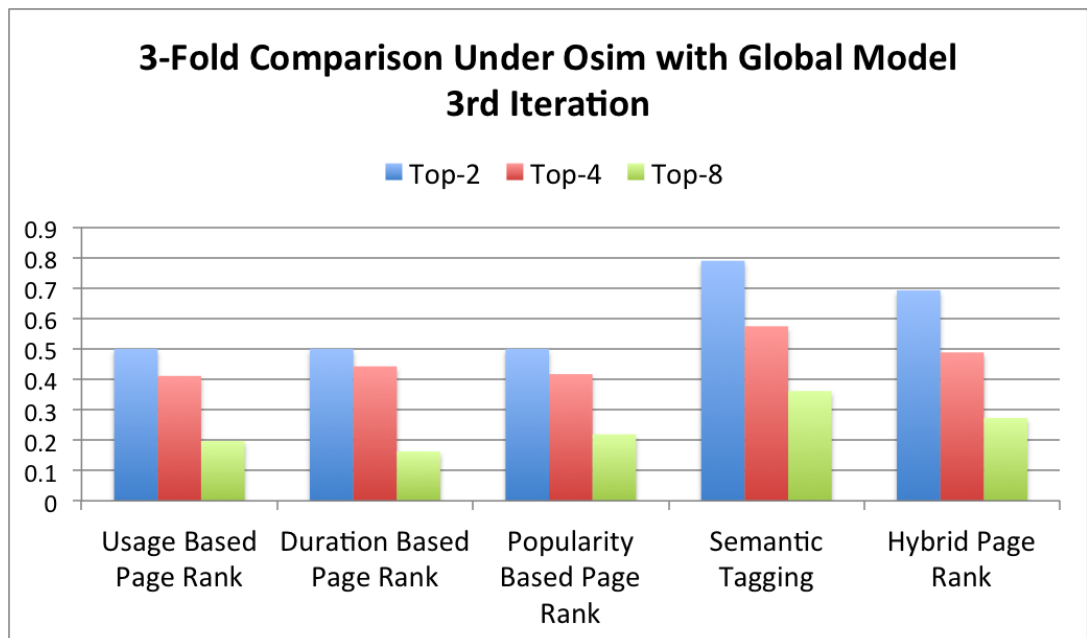


Figure B.11: Third Iteration of 3-Fold Validation with Global Model Under Osim Similarity



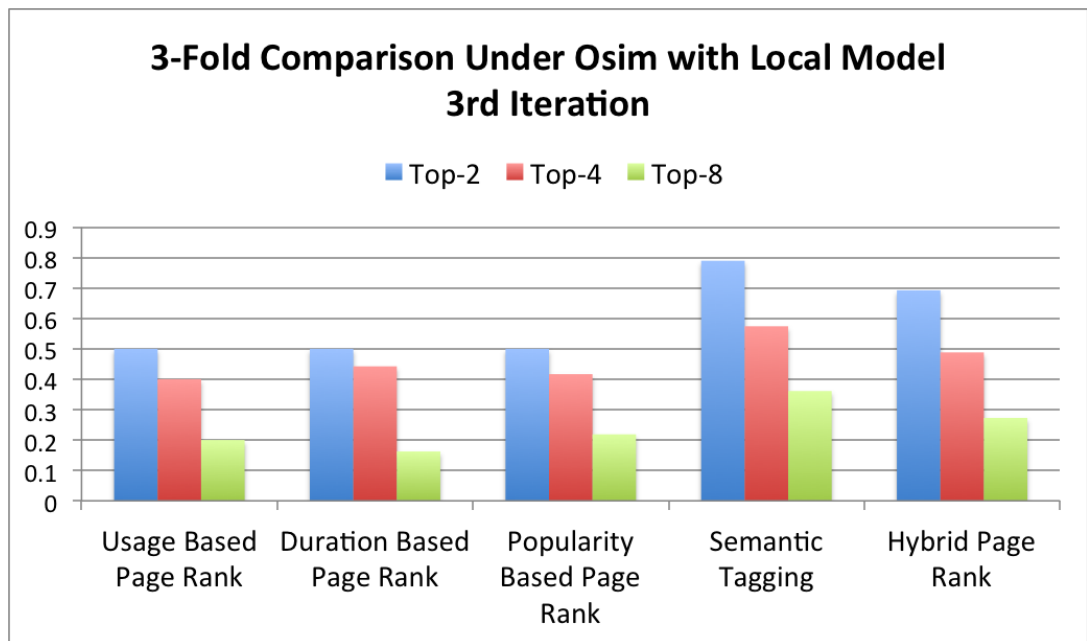


Figure B.12: Third Iteration of 3-Fold Validation with Local Model Under Osim Similarity

**B.2 5-Fold Cross Validation Detailed Results**

**B.2.1 Ksim Similarity Measures**

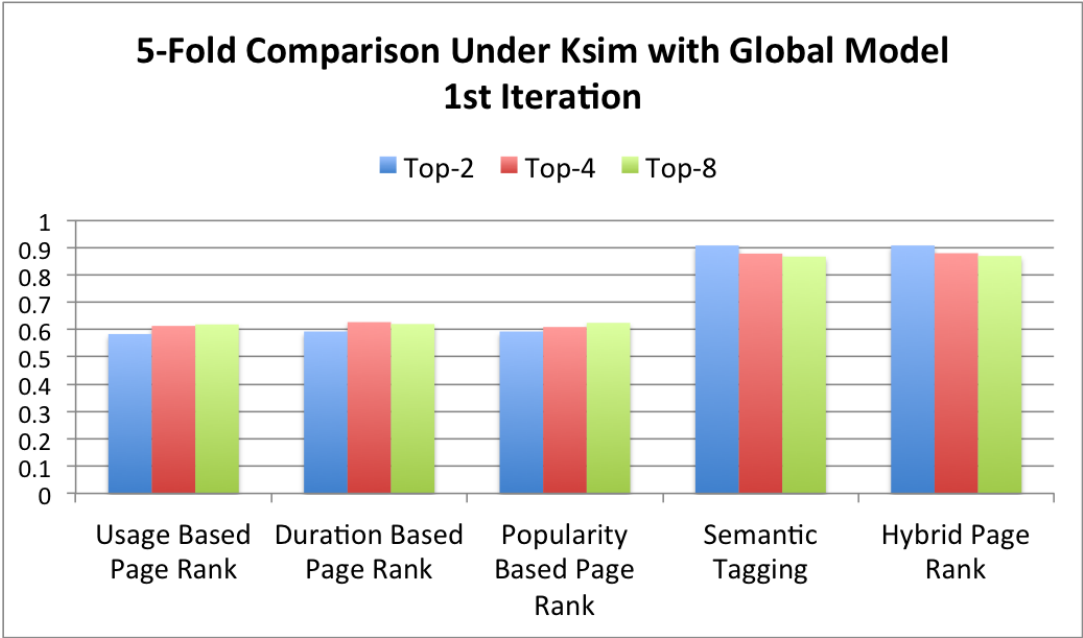


Figure B.13: First Iteration of 5-Fold Validation with Global Model Under Ksim Similarity

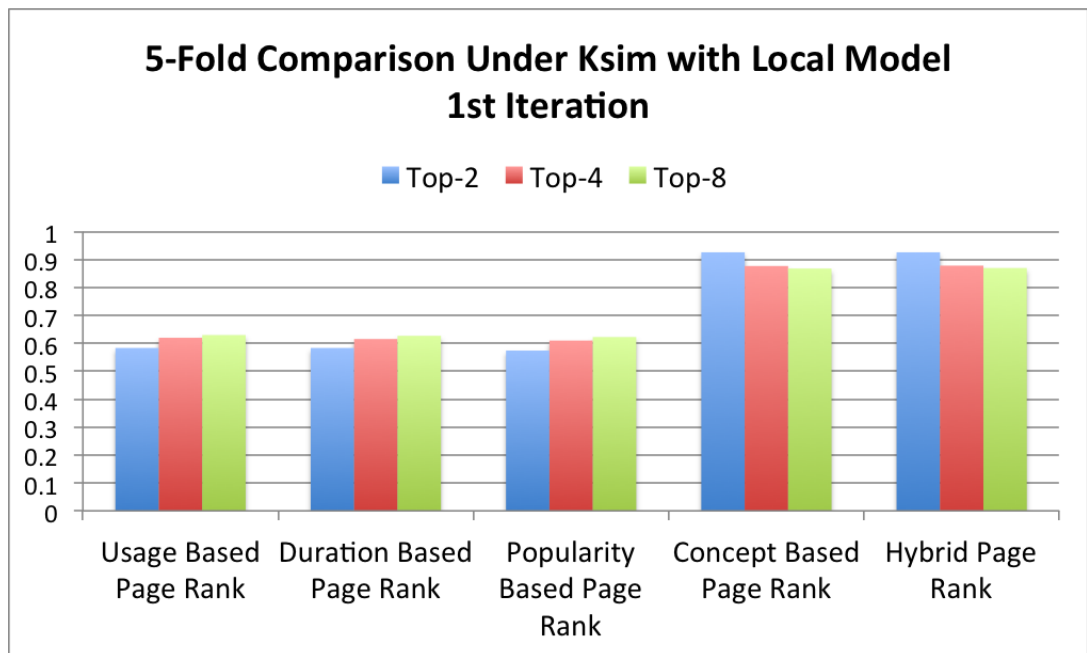


Figure B.14: First Iteration of 5-Fold Validation with Local Model Under Ksim Similarity

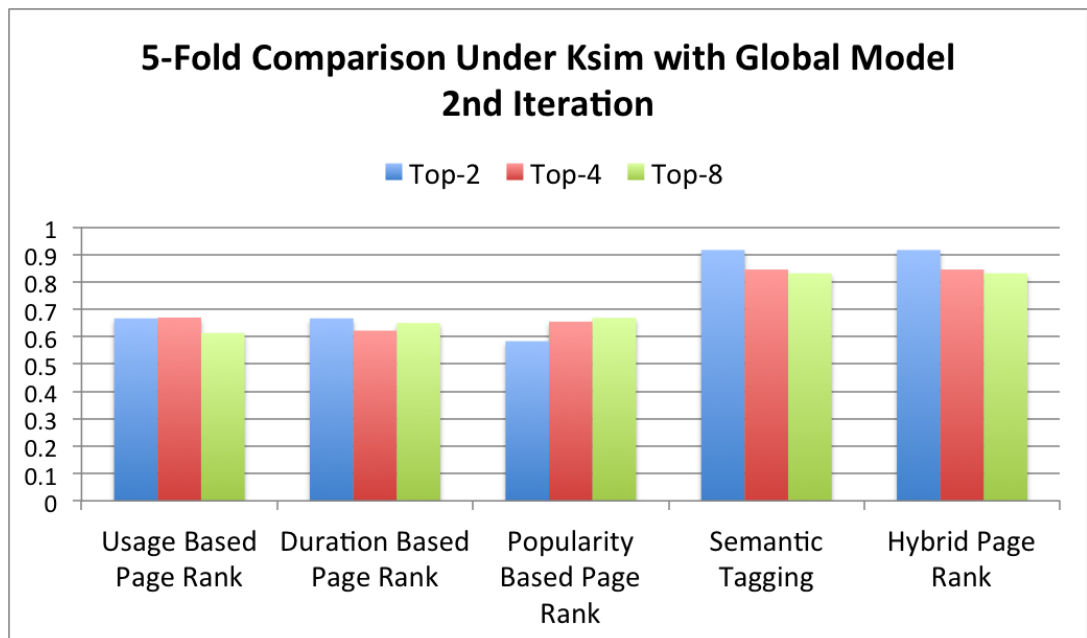


Figure B.15: Second Iteration of 5-Fold Validation with Global Model Under Ksim Similarity

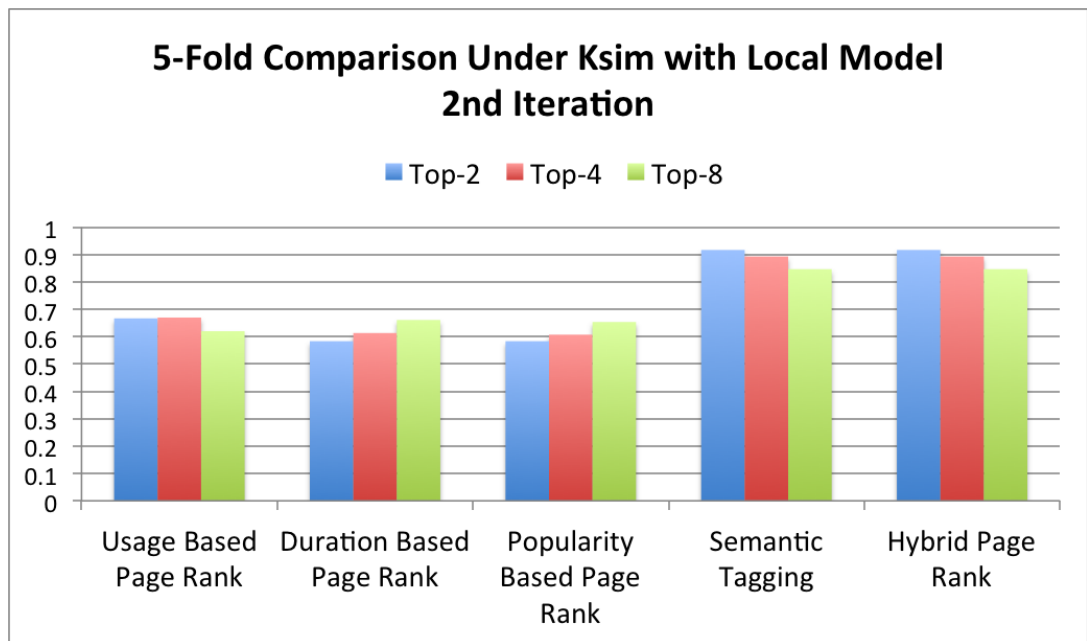


Figure B.16: Second Iteration of 5-Fold Validation with Local Model Under Ksim Similarity

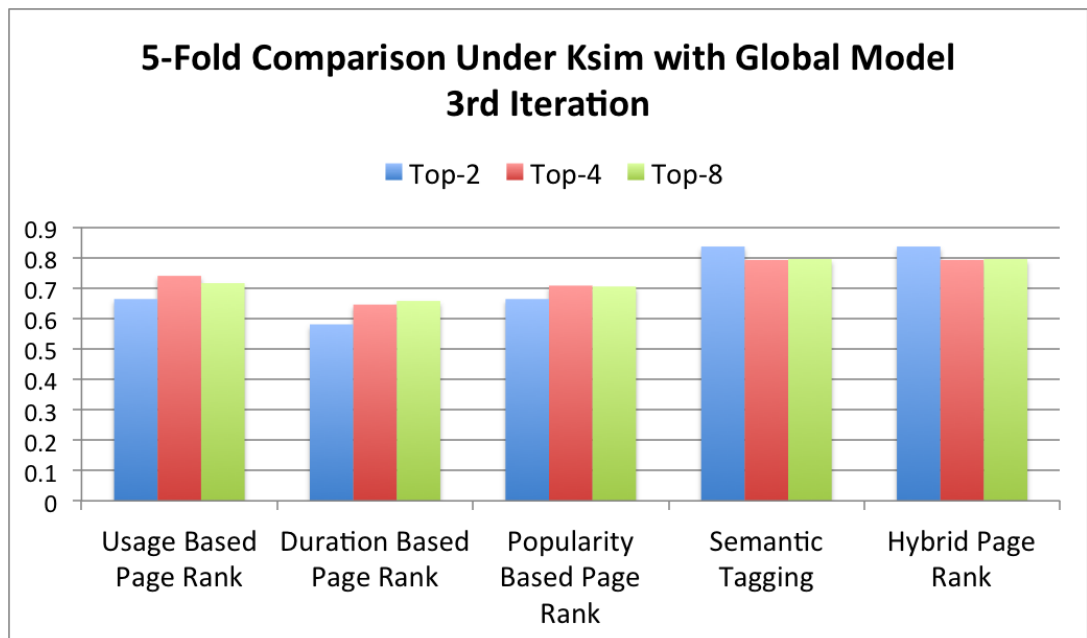


Figure B.17: Third Iteration of 5-Fold Validation with Global Model Under Ksim Similarity

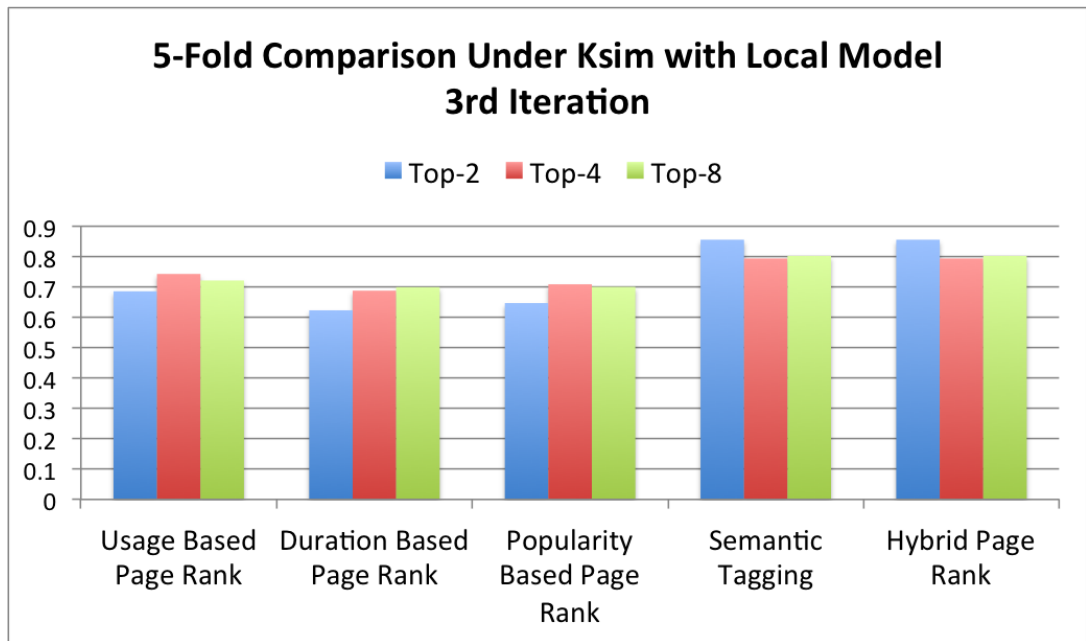


Figure B.18: Third Iteration of 5-Fold Validation with Local Model Under Ksim Similarity

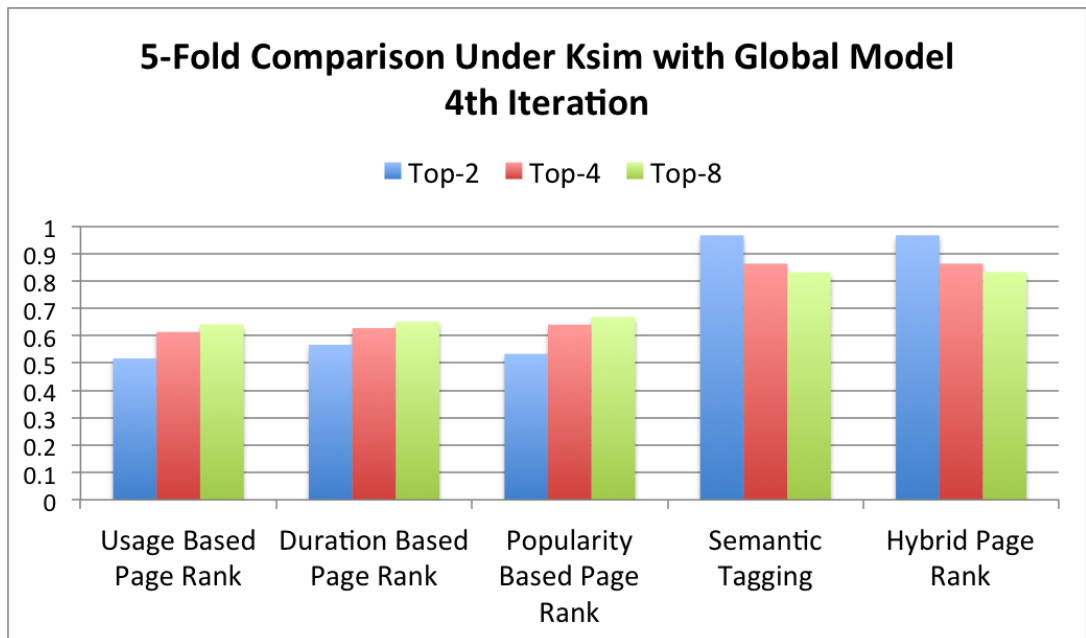


Figure B.19: Fourth Iteration of 5-Fold Validation with Global Model Under Ksim Similarity

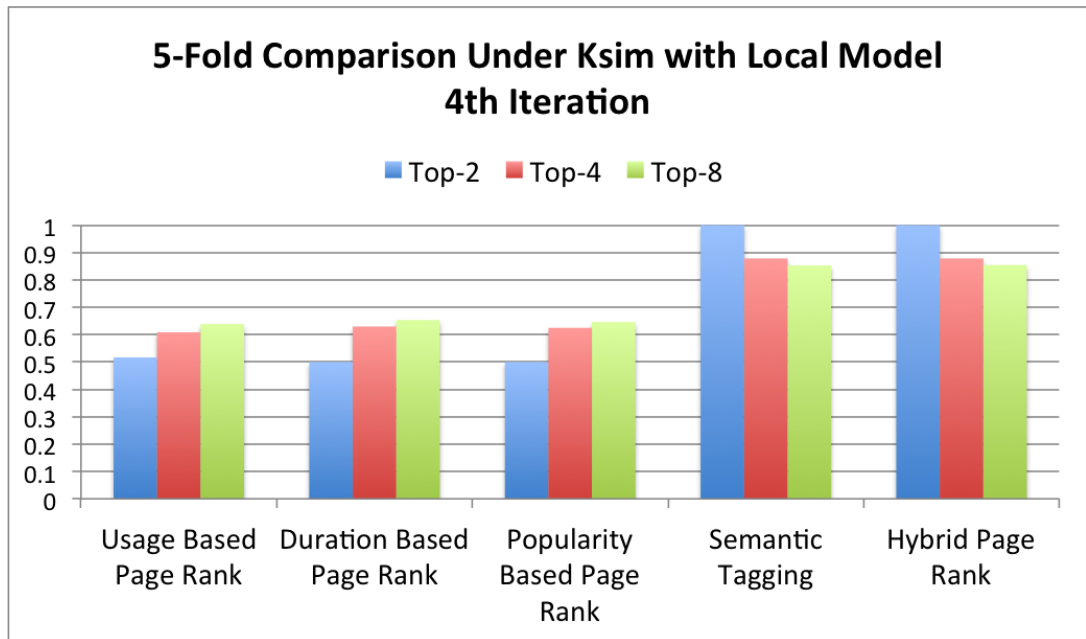


Figure B.20: Fourth Iteration of 5-Fold Validation with Local Model Under Ksim Similarity

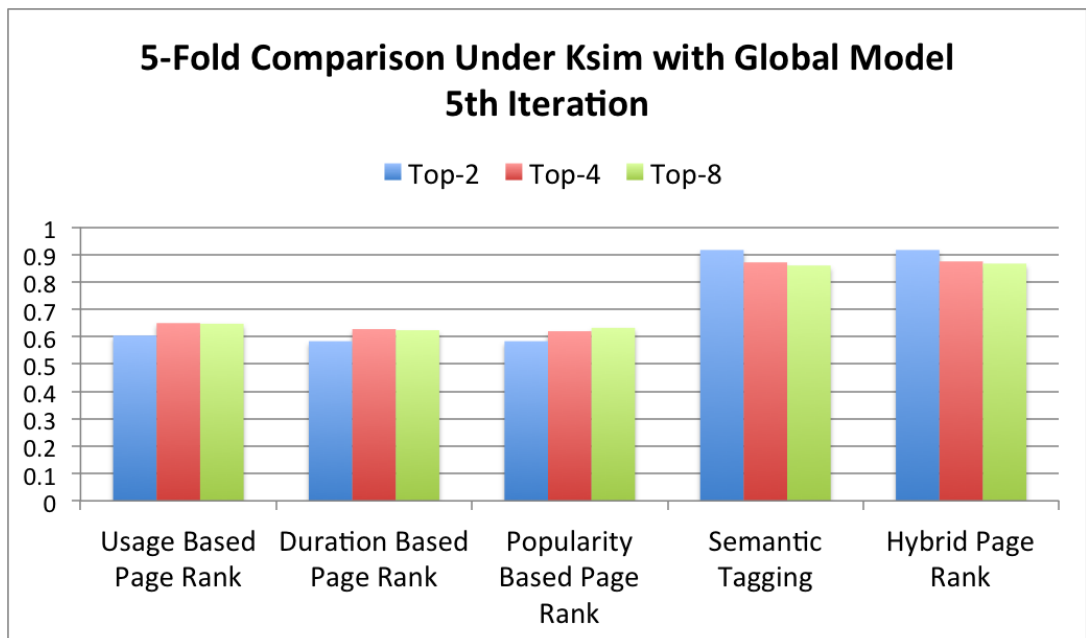


Figure B.21: Fifth Iteration of 5-Fold Validation with Global Model Under Ksim Similarity

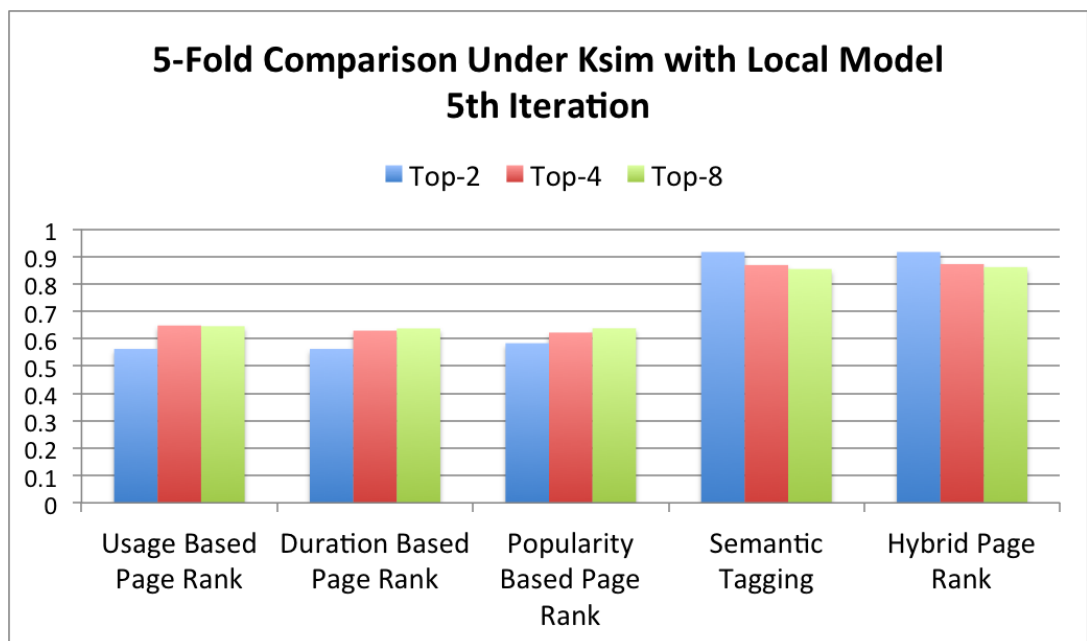


Figure B.22: Fifth Iteration of 5-Fold Validation with Local Model Under Ksim Similarity

**B.2.2 Osim Similarity Measures**

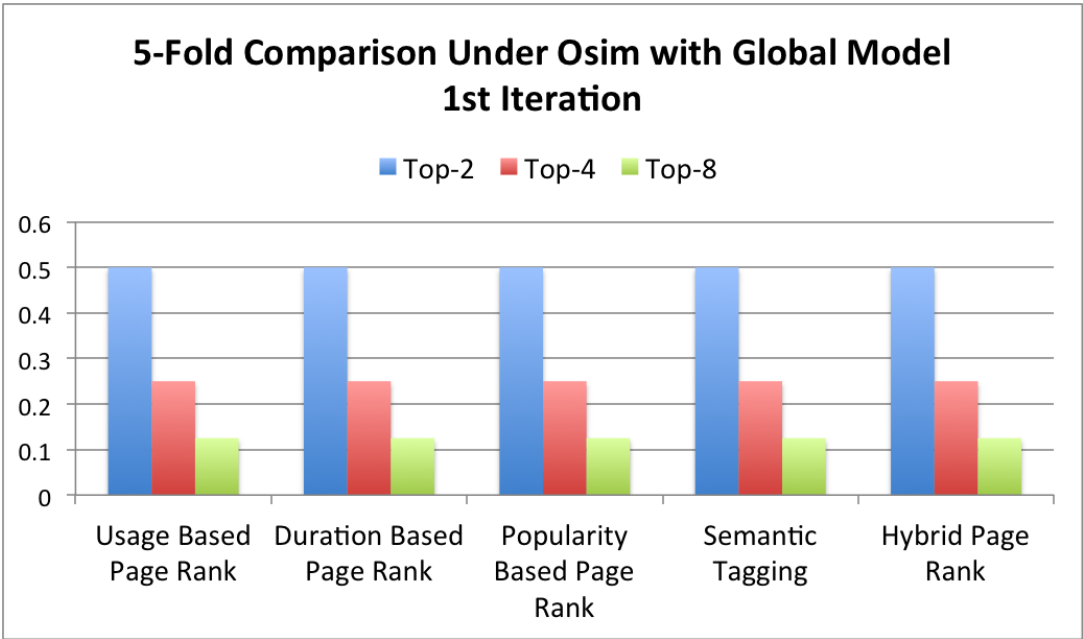


Figure B.23: First Iteration of 5-Fold Validation with Global Model Under Osim Similarity



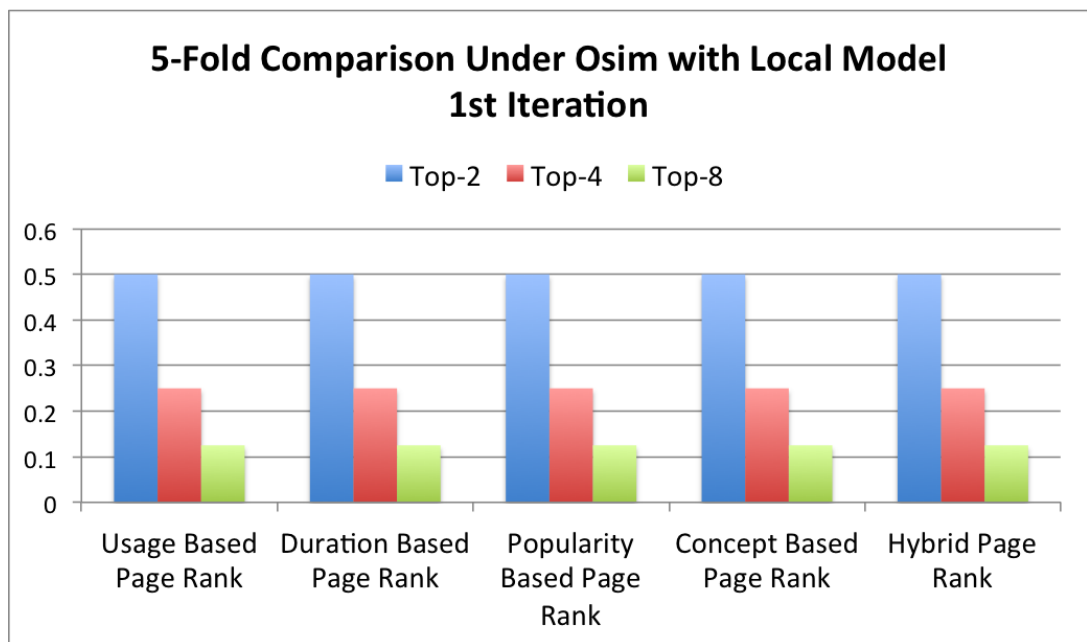


Figure B.24: First Iteration of 5-Fold Validation with Local Model Under Osim Similarity

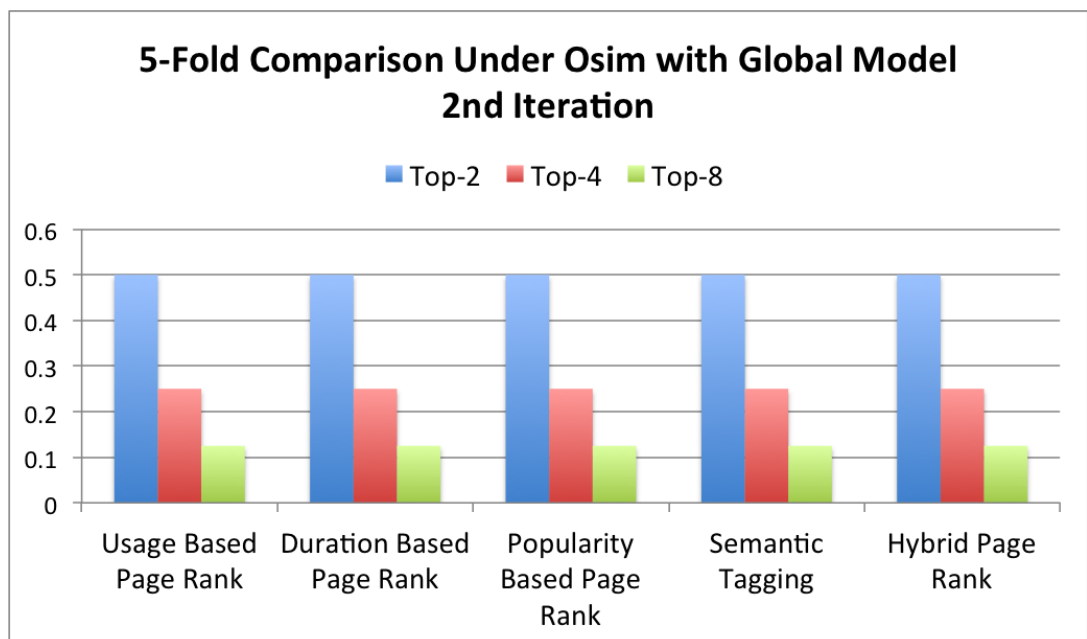


Figure B.25: Second Iteration of 5-Fold Validation with Global Model Under Osim Similarity

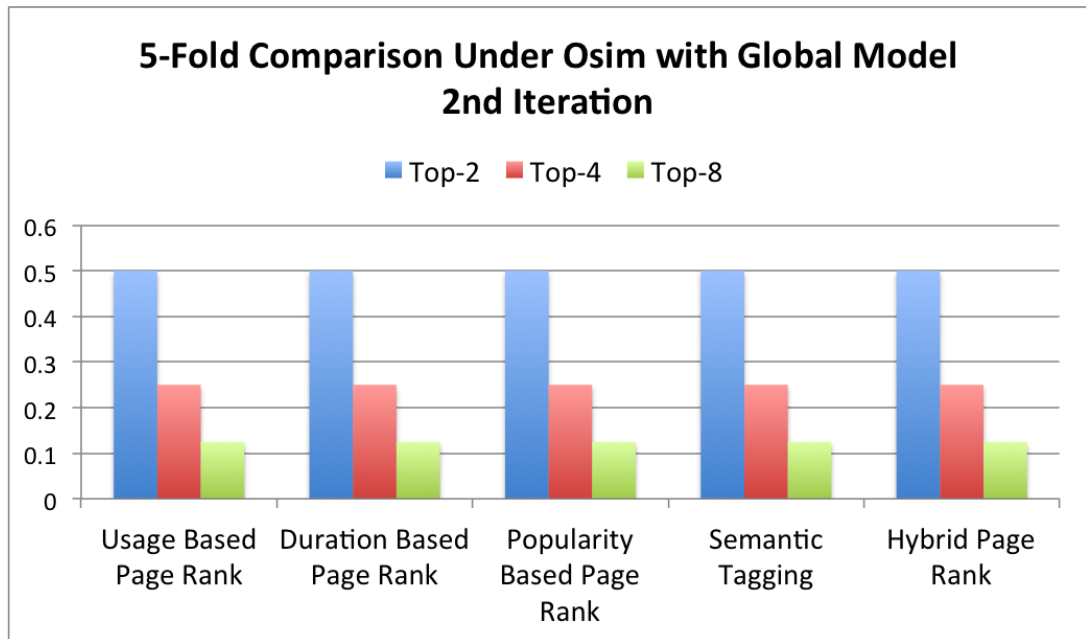


Figure B.26: Second Iteration of 5-Fold Validation with Local Model Under Osim Similarity

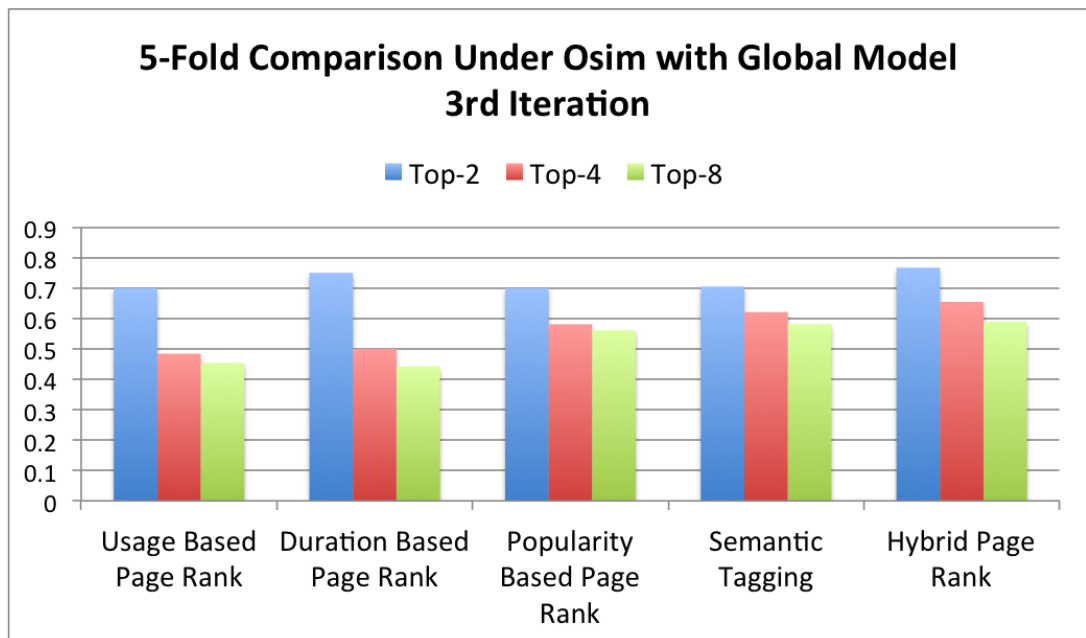


Figure B.27: Third Iteration of 5-Fold Validation with Global Model Under Osim Similarity

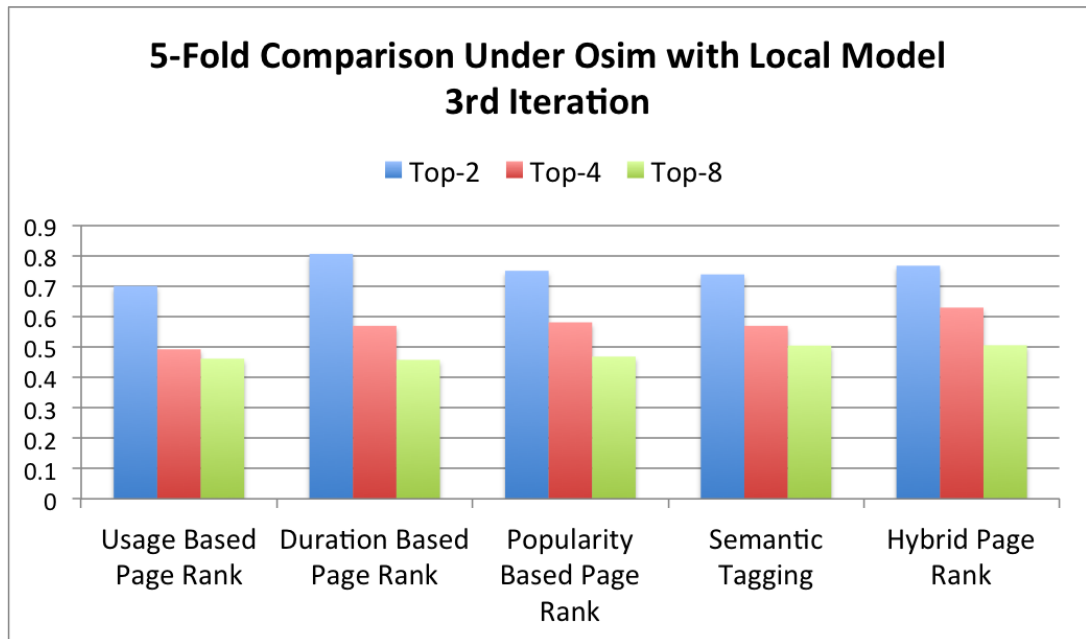


Figure B.28: Third Iteration of 5-Fold Validation with Local Model Under Osim Similarity

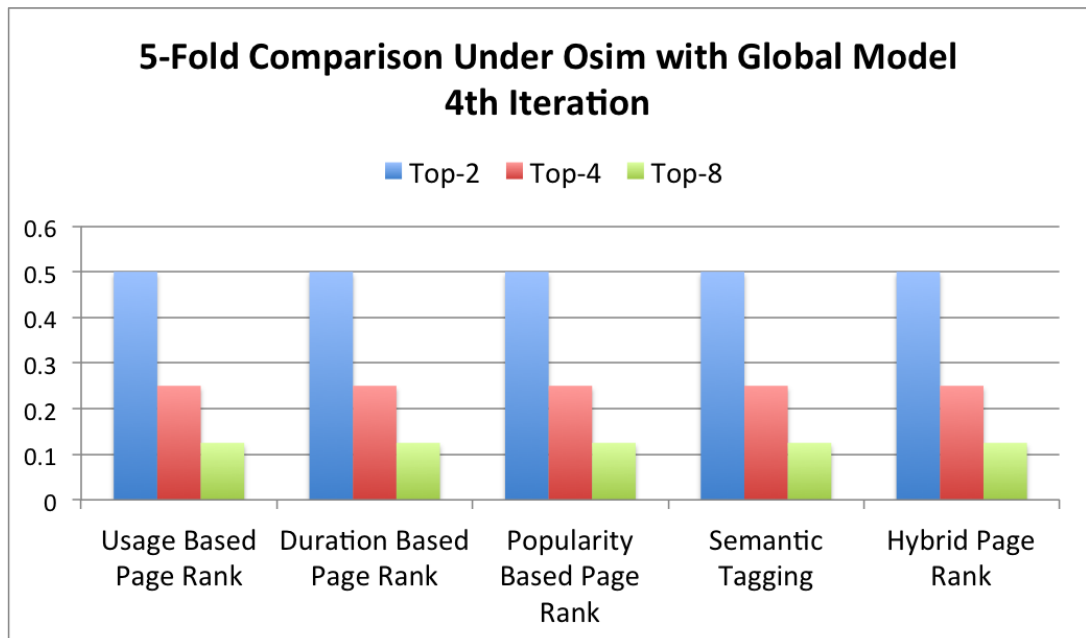


Figure B.29: Fourth Iteration of 5-Fold Validation with Global Model Under Osim Similarity

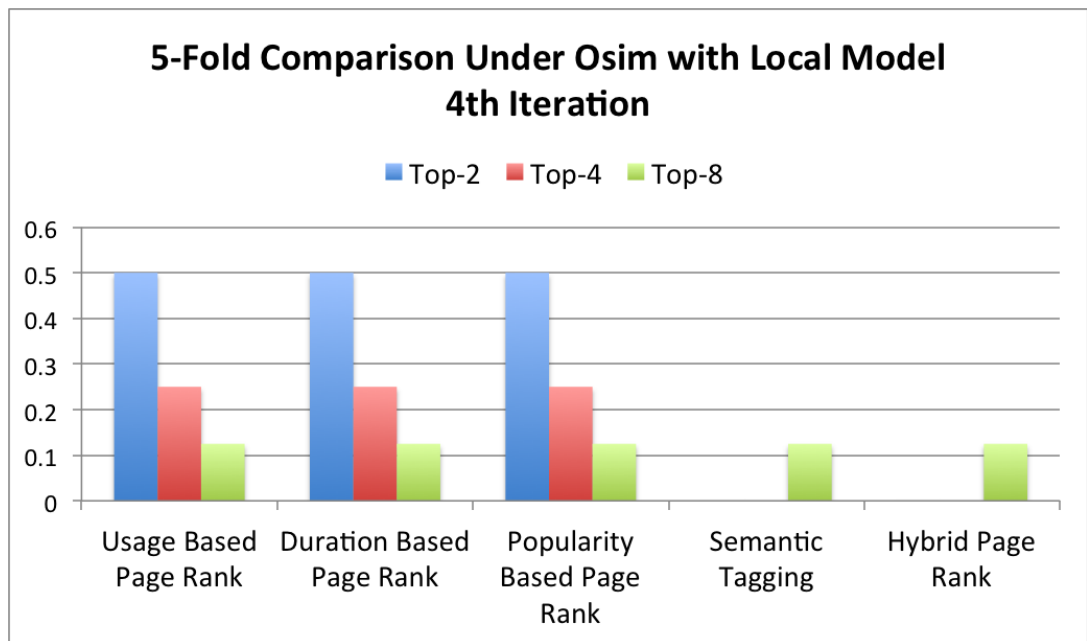


Figure B.30: Fourth Iteration of 5-Fold Validation with Local Model Under Osim Similarity

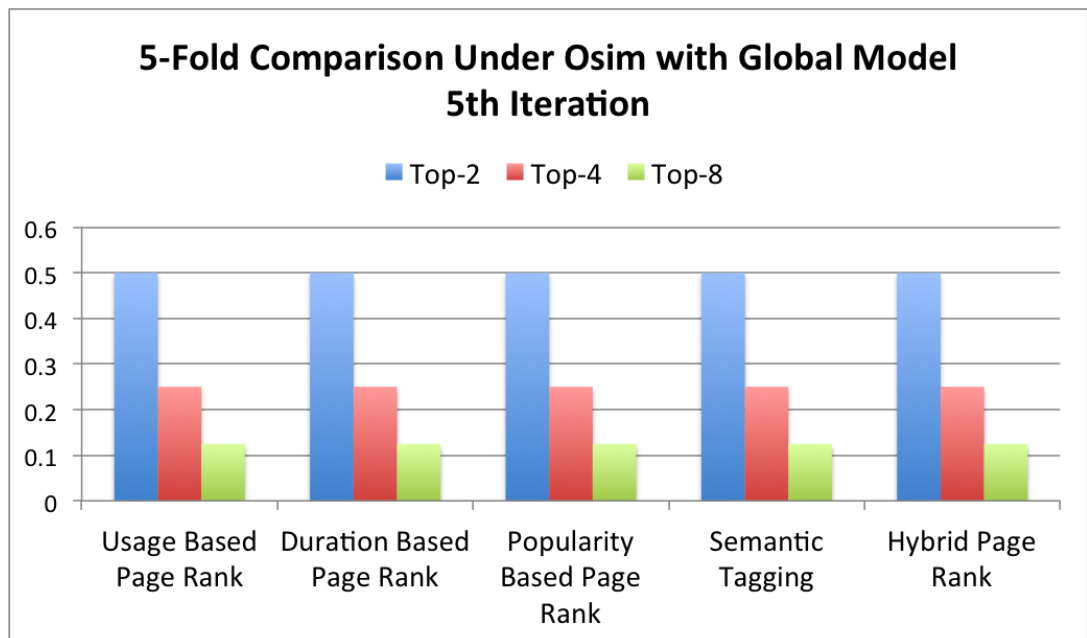


Figure B.31: Fifth Iteration of 5-Fold Validation with Global Model Under Osim Similarity

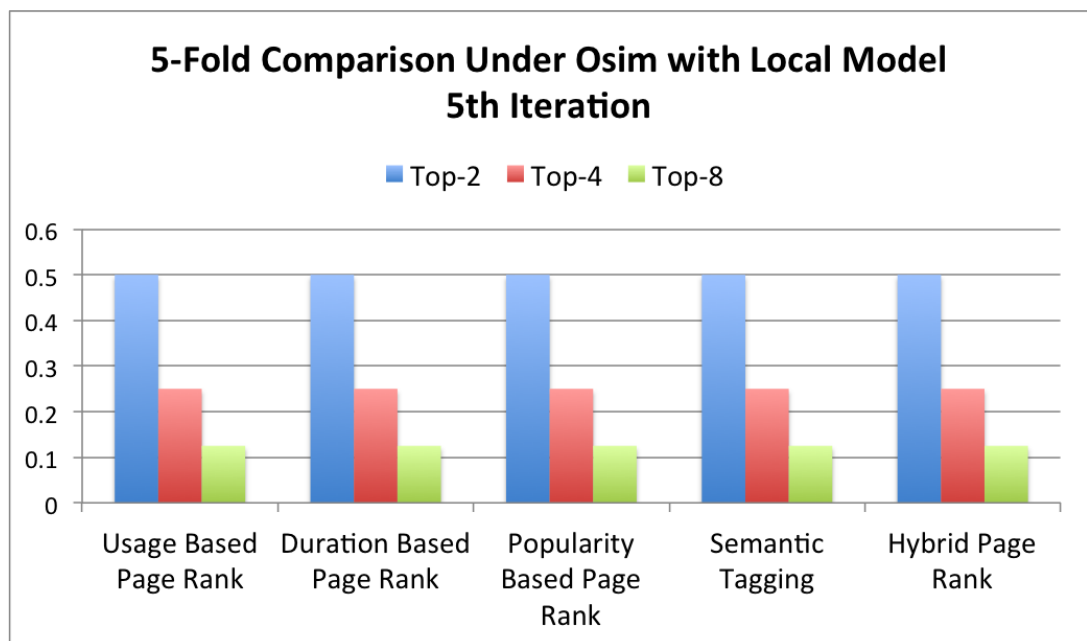


Figure B.32: Fifth Iteration of 5-Fold Validation with Local Model Under Osim Similarity

### B.3 10 Fold Cross Validation Detailed Results

In 10-fold cross validation, because of the variety of the unique page number is very high in the data set, in some iterations all of the methods do not produce next page predictions. For this reason the results is given in a summary format with all folds together for each method with global, local model in *Ksim* and *Osim* metrics.

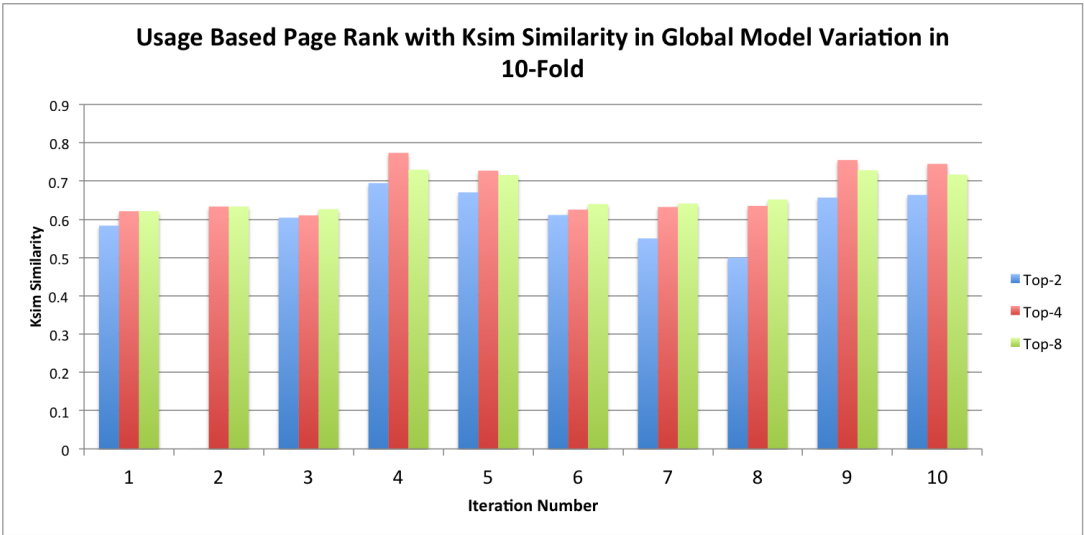


Figure B.33: Usage Based Page Rank 10-Fold Validation with Global Model Under Ksim Similarity

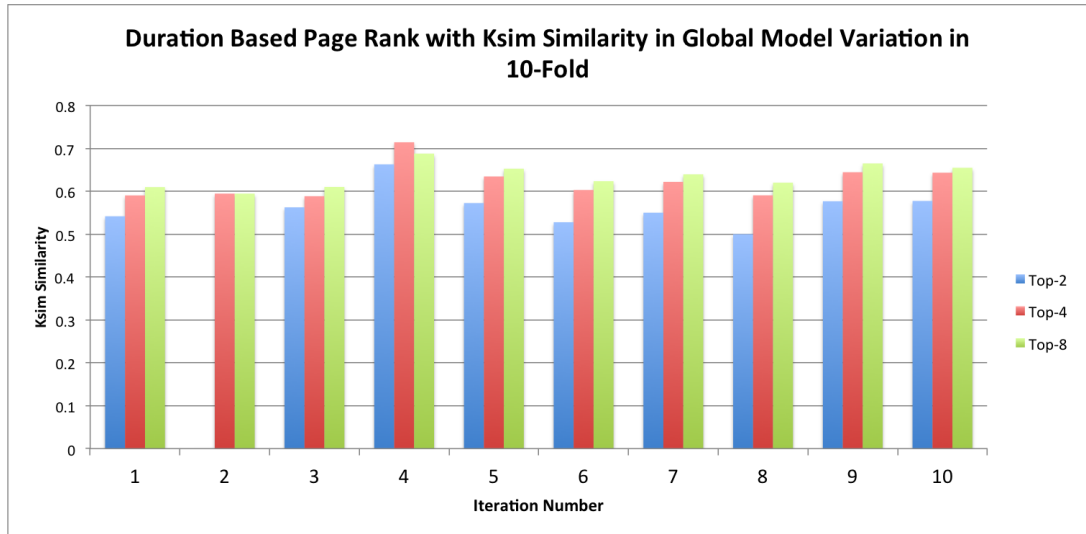


Figure B.34: Duration Based Page Rank 10-Fold Validation with Global Model Under Ksim Similarity

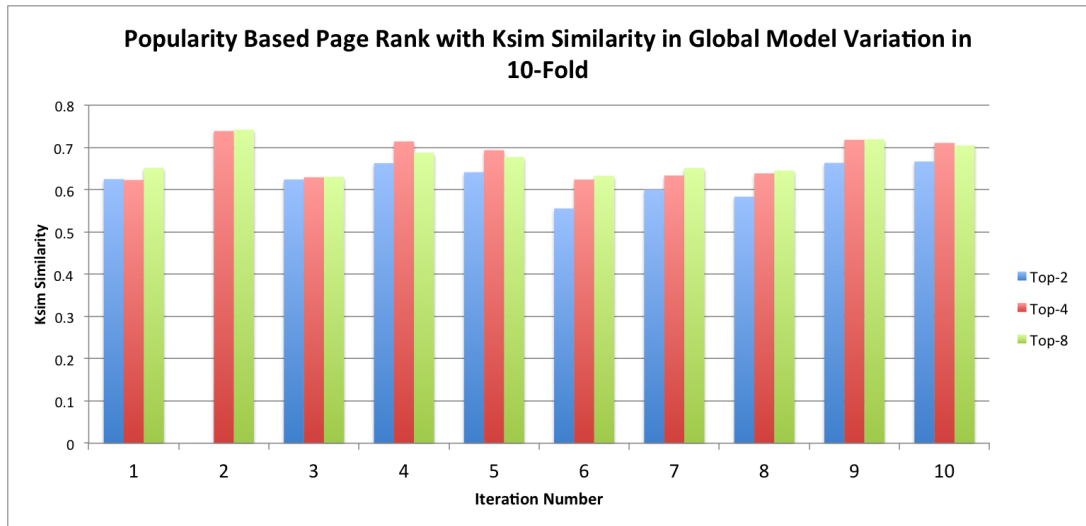


Figure B.35: Popularity Based Page Rank 10-Fold Validation with Global Model Under Ksim Similarity

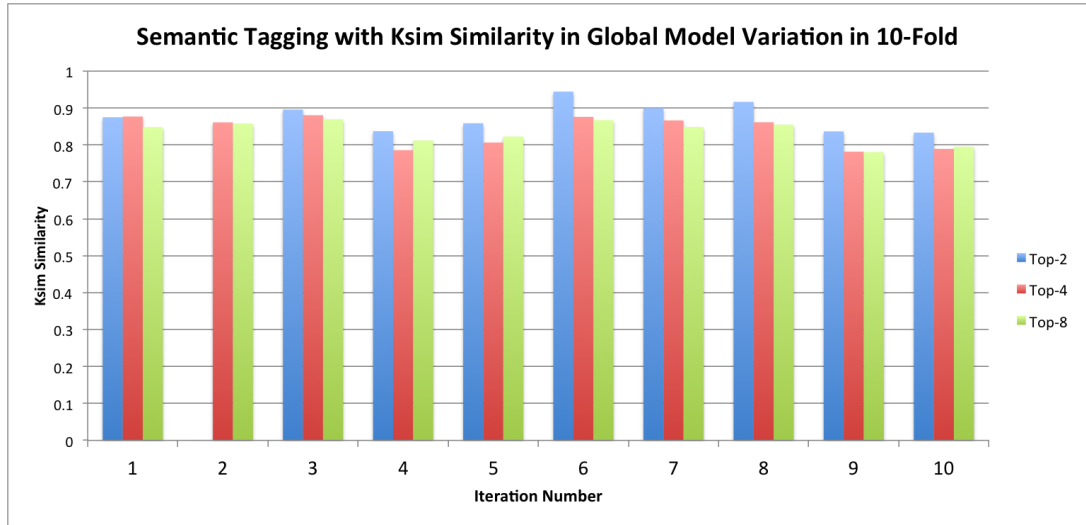


Figure B.36: Semantic Tagging 10-Fold Validation with Global Model Under Ksim Similarity

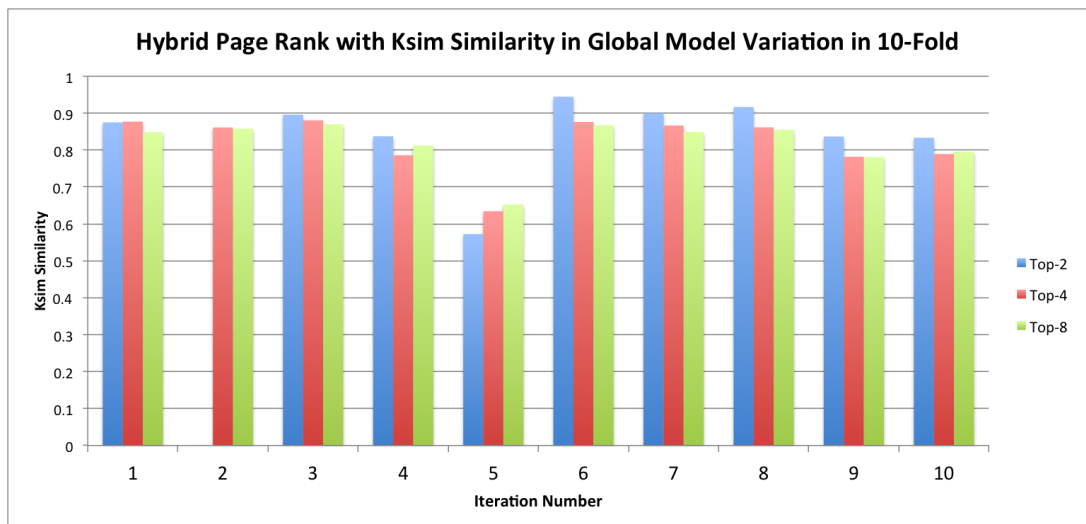


Figure B.37: Hybrid Page Rank 10-Fold Validation with Global Model Under Ksim Similarity



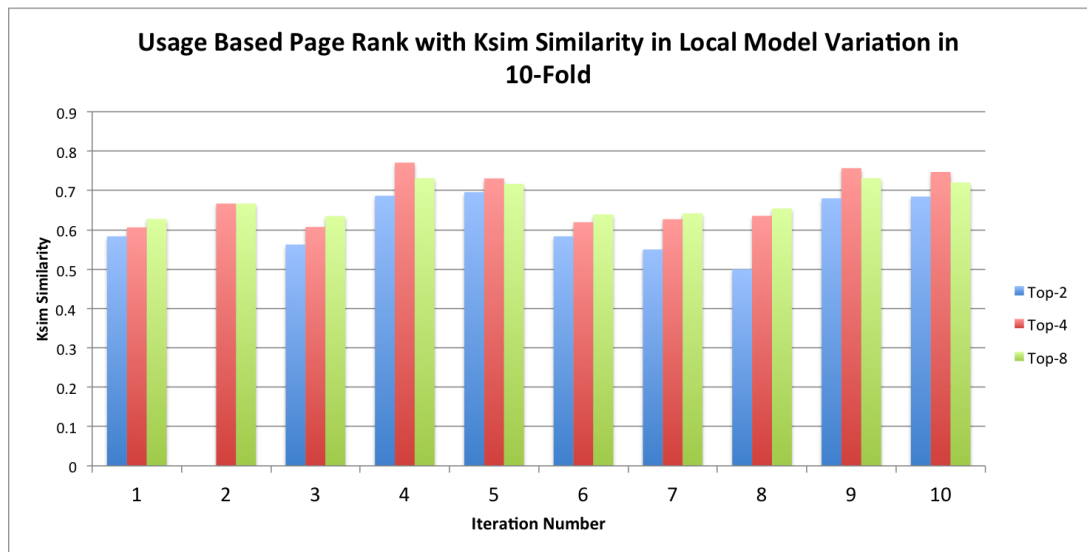


Figure B.38: Usage Based Page Rank 10-Fold Validation with Local Model Under Ksim Similarity

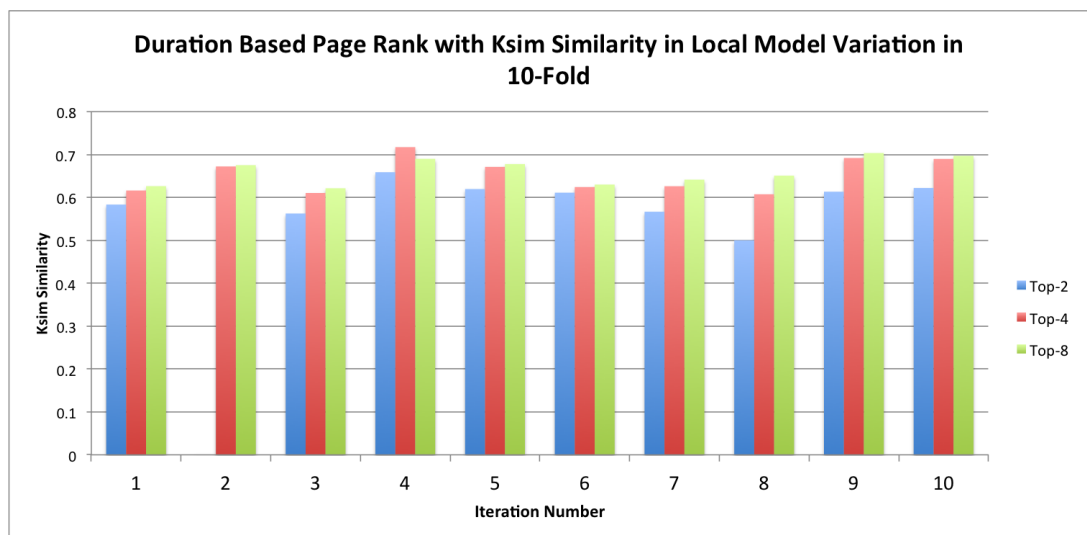


Figure B.39: Duration Based Page Rank 10-Fold Validation with Local Model Under Ksim Similarity

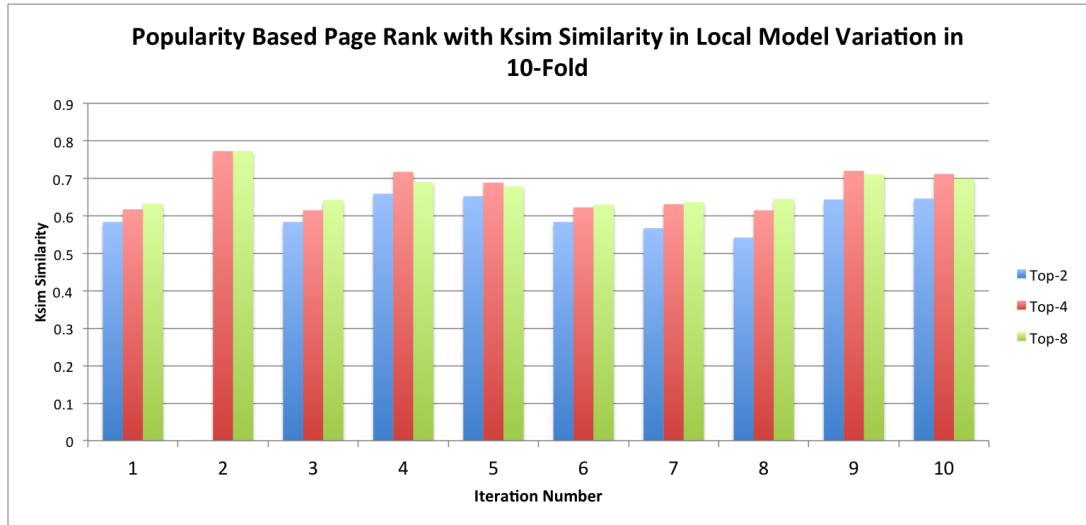


Figure B.40: Popularity Based Page Rank 10-Fold Validation with Local Model Under Ksim Similarity

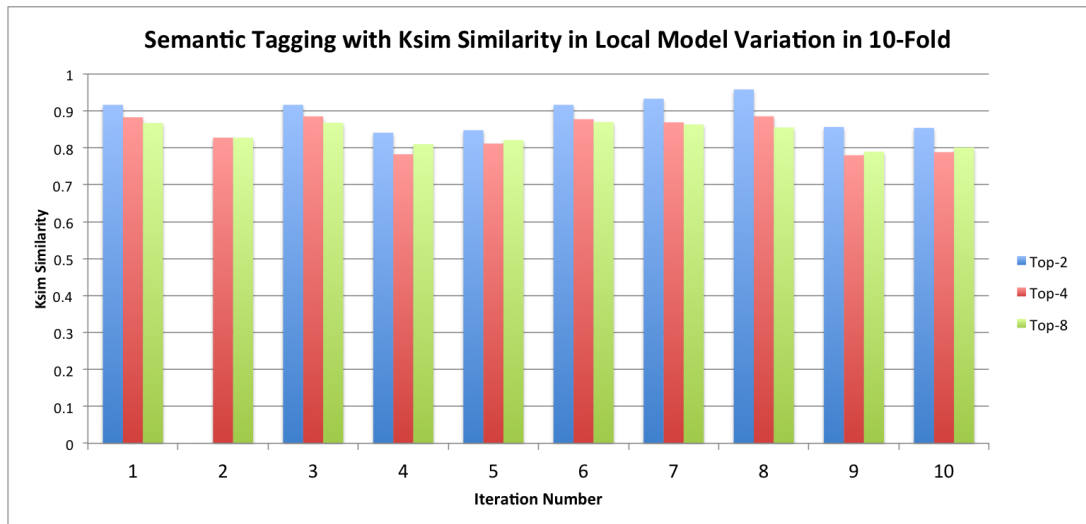


Figure B.41: Semantic Tagging 10-Fold Validation with Local Model Under Ksim Similarity

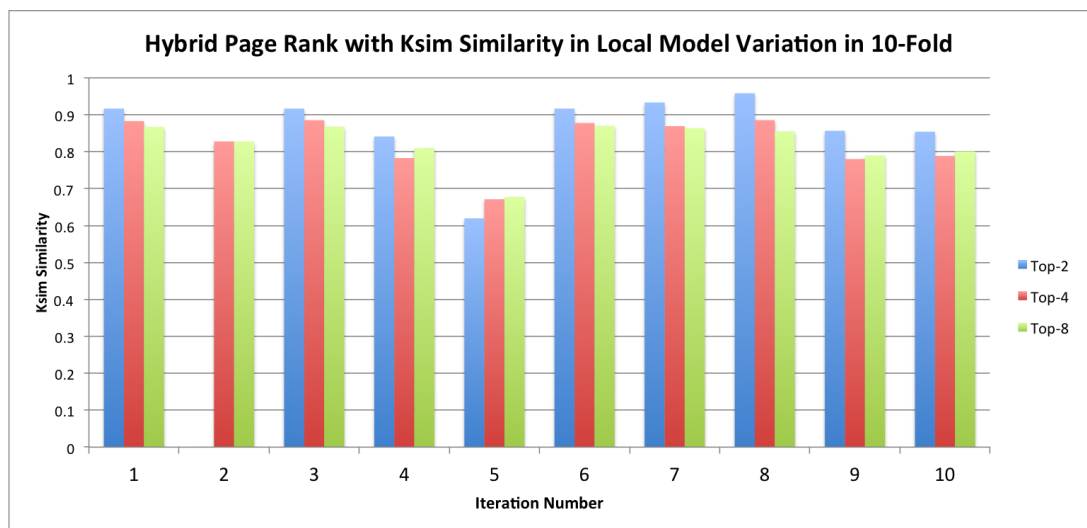


Figure B.42: Hybrid Page Rank 10-Fold Validation with Local Model Under Ksim Similarity

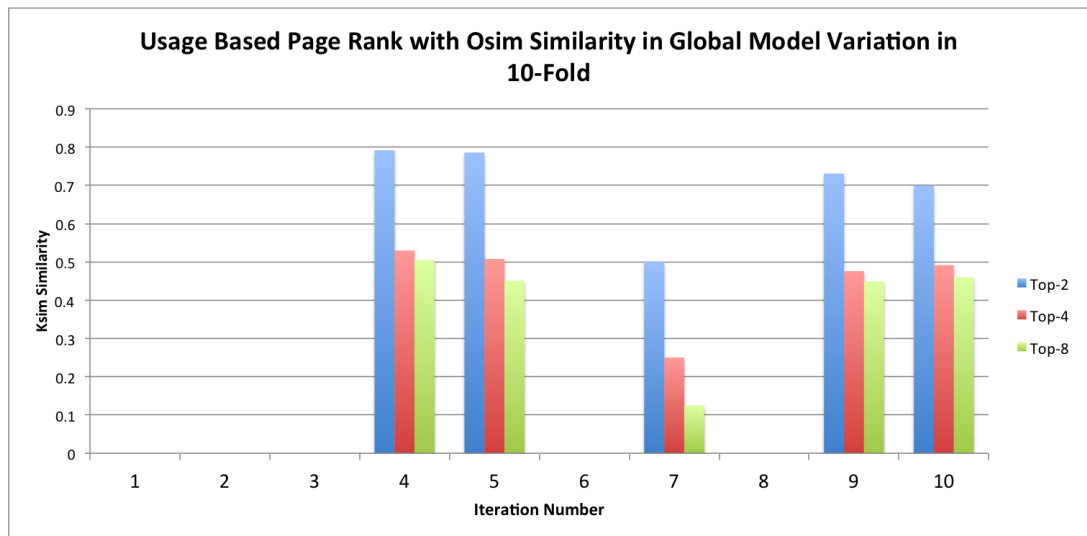


Figure B.43: Usage Based Page Rank 10-Fold Validation with Global Model Under Osim Similarity

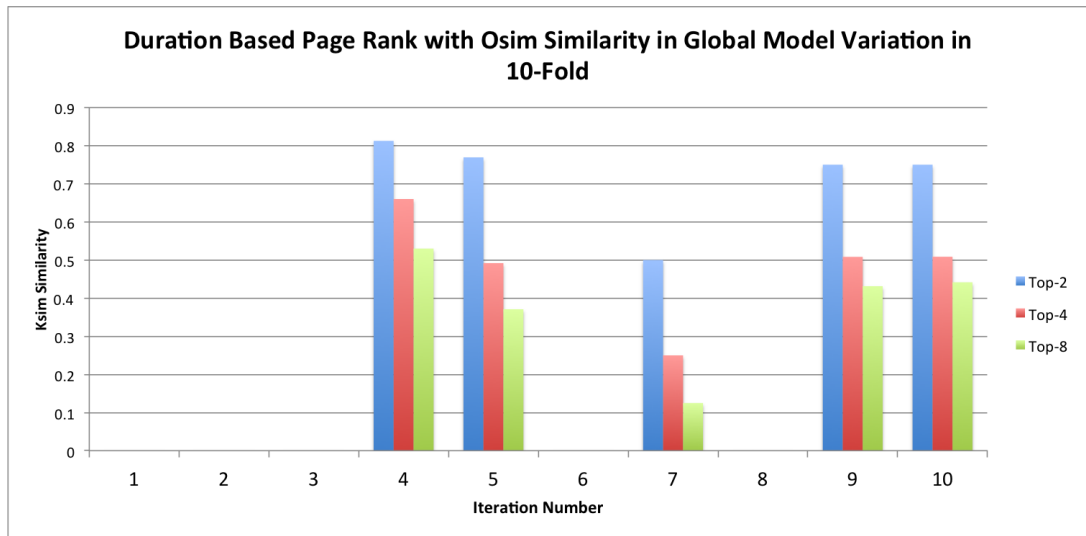


Figure B.44: Duration Based Page Rank 10-Fold Validation with Global Model Under Osim Similarity

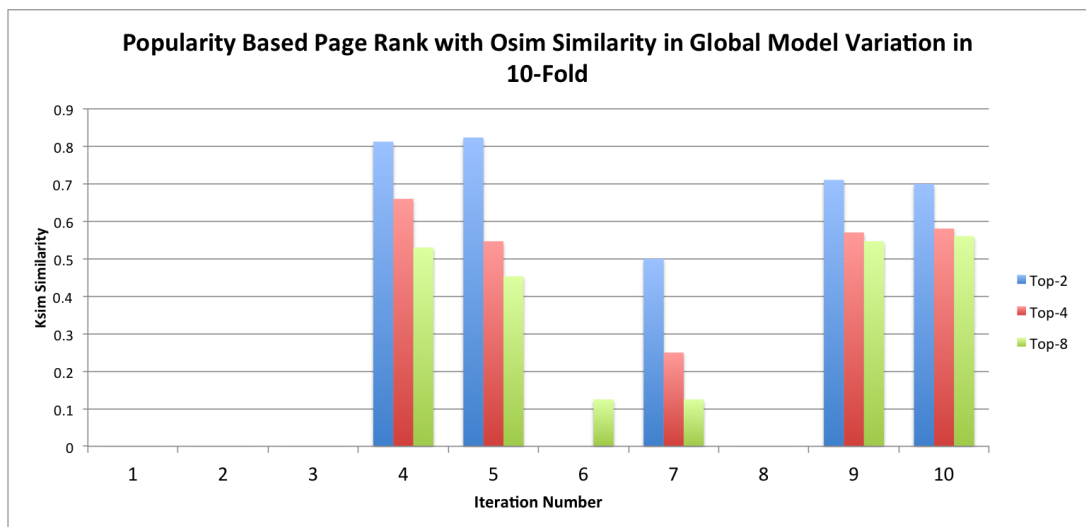


Figure B.45: Popularity Based Page Rank 10-Fold Validation with Global Model Under Osim Similarity

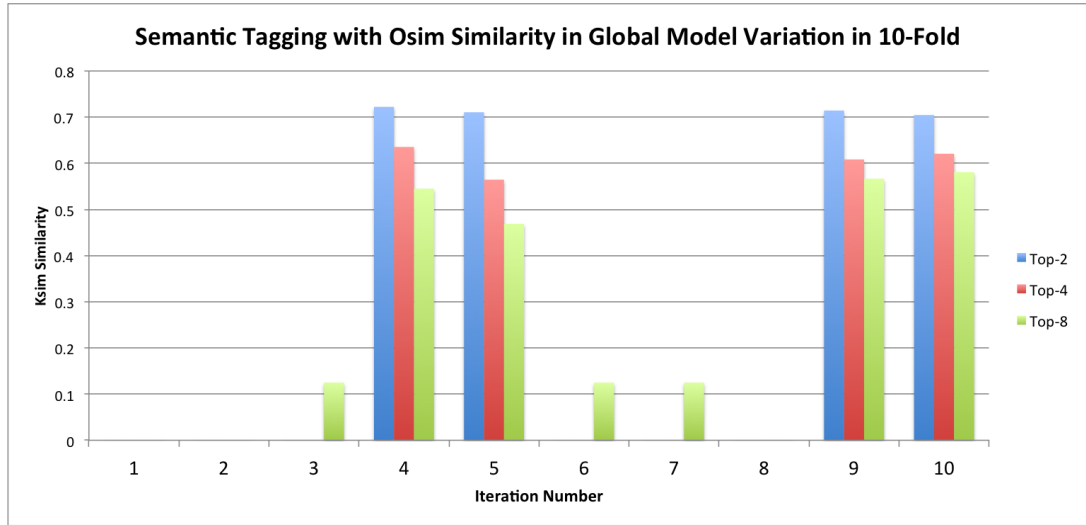


Figure B.46: Semantic Tagging 10-Fold Validation with Global Model Under Osim Similarity

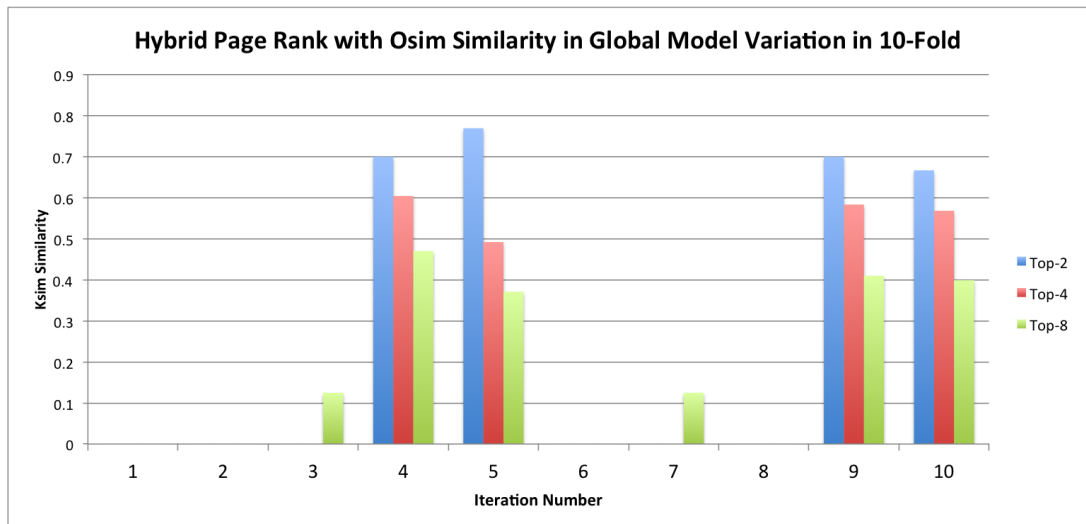


Figure B.47: Hybrid Page Rank 10-Fold Validation with Global Model Under Osim Similarity

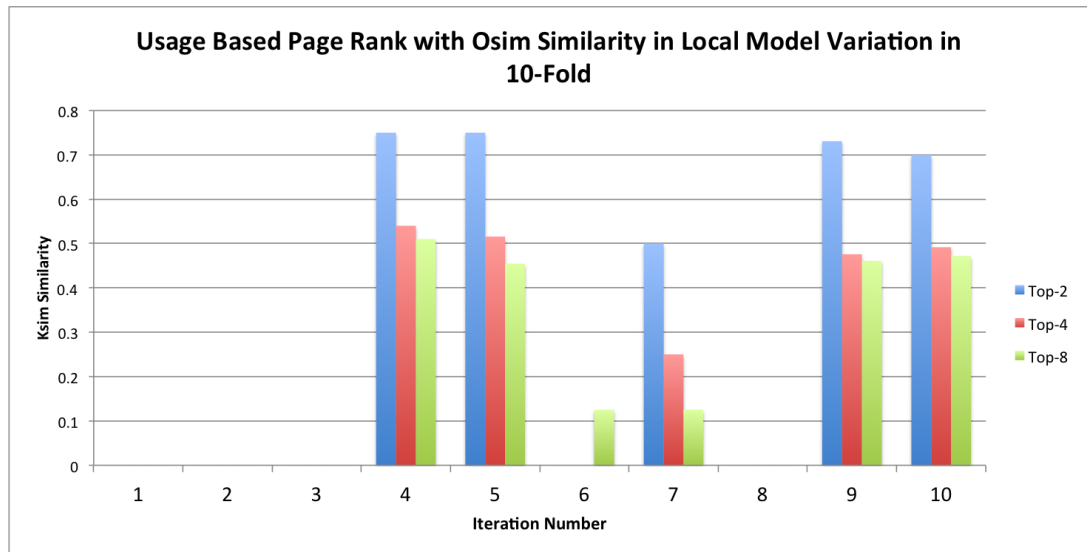


Figure B.48: Usage Based Page Rank 10-Fold Validation with Local Model Under Osim Similarity

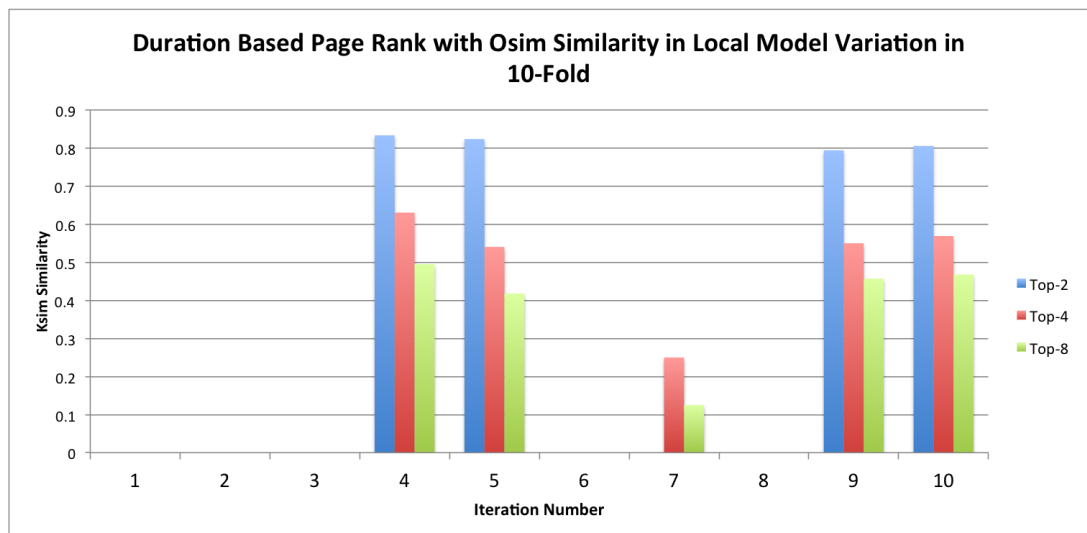


Figure B.49: Duration Based Page Rank 10-Fold Validation with Local Model Under Osim Similarity

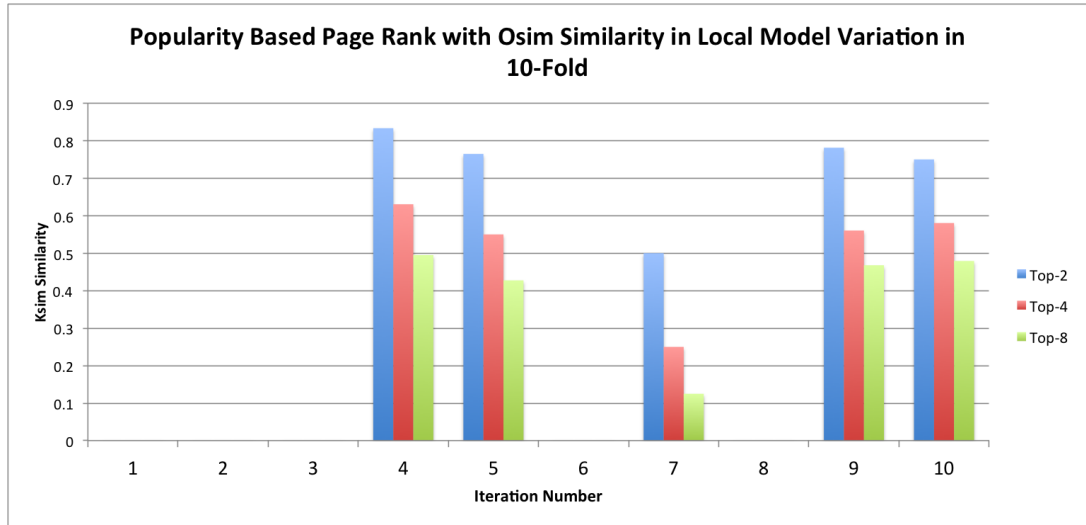


Figure B.50: Popularity Based Page Rank 10-Fold Validation with Local Model Under Osim Similarity

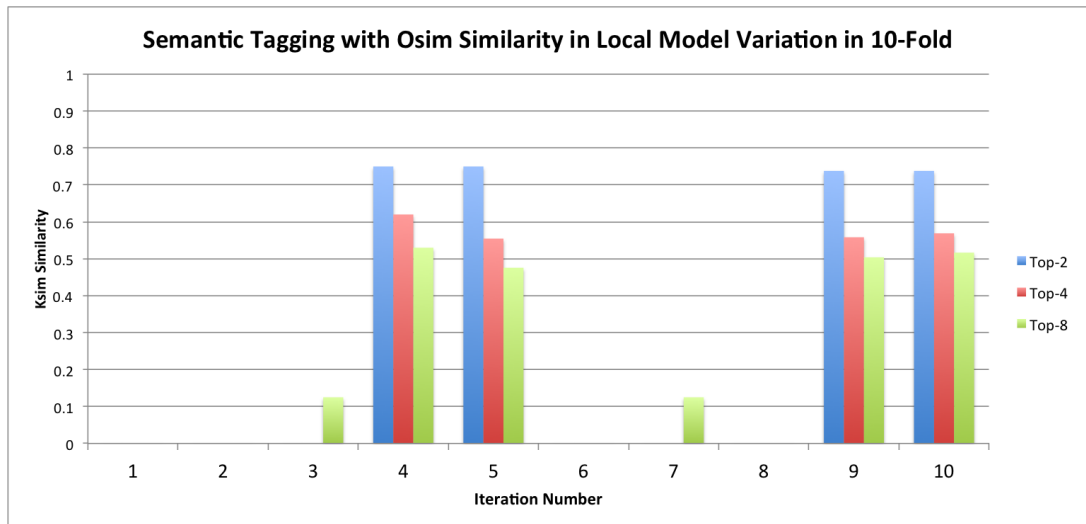


Figure B.51: Semantic Tagging 10-Fold Validation with Local Model Under Osim Similarity



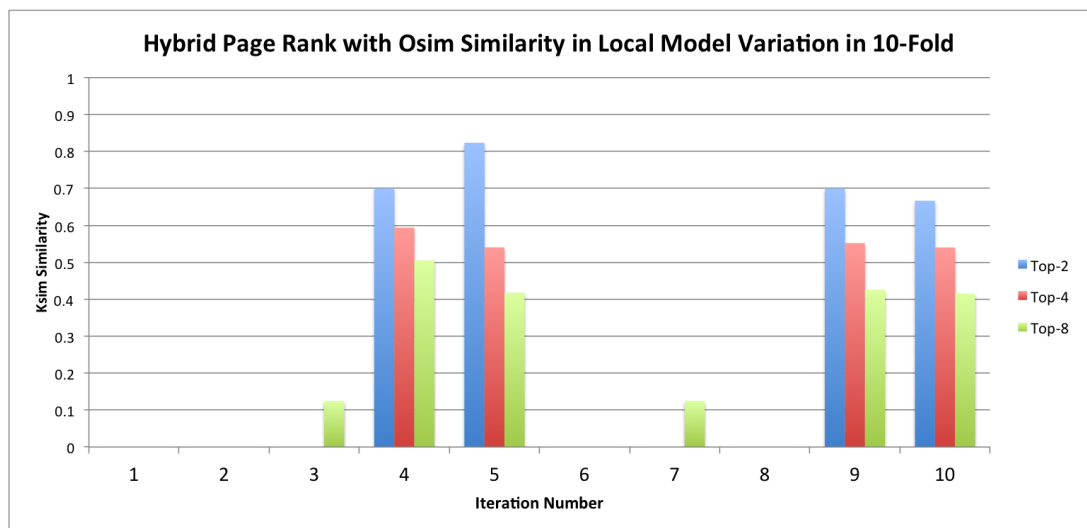


Figure B.52: Hybrid Page Rank 10-Fold Validation with Local Model Under Osim Similarity

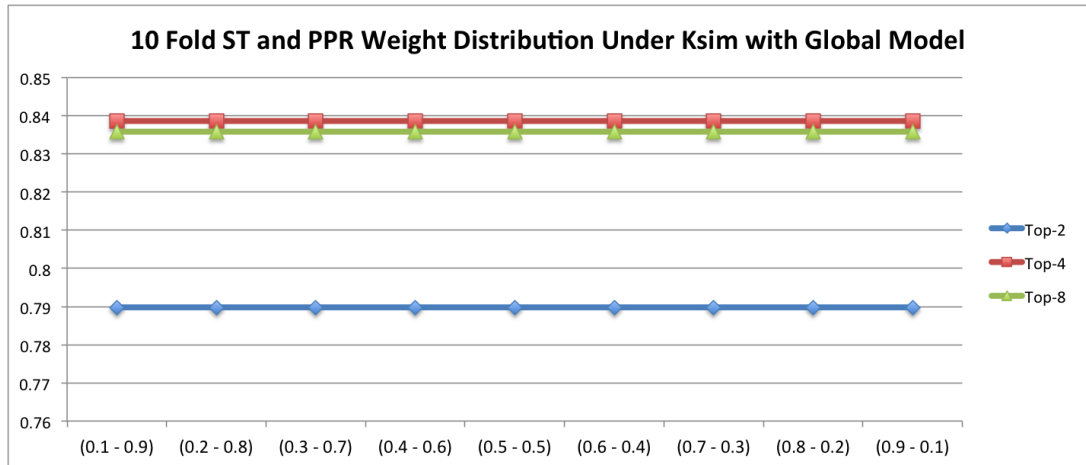


Figure B.53: 10-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Global Model

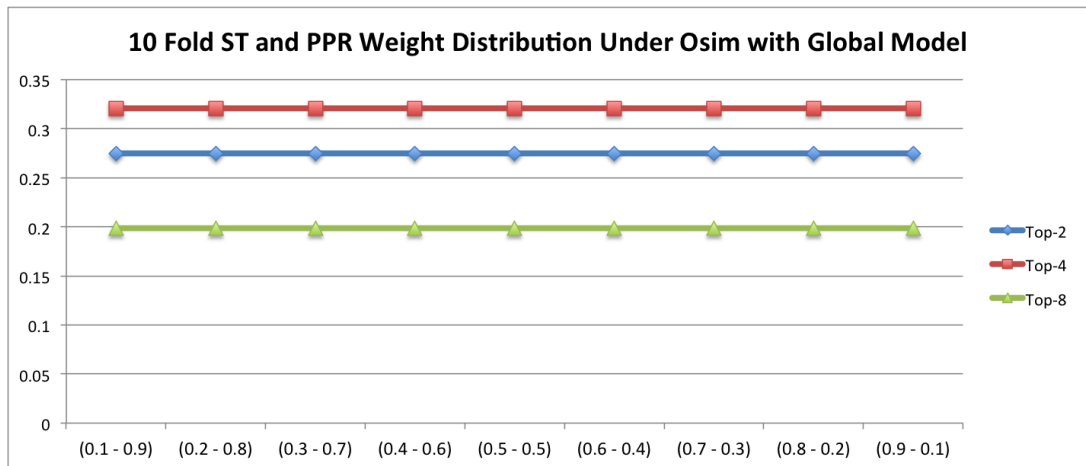


Figure B.54: 10-Fold Validation ST and PPR Weight Effects on HPR under Osim with Global Model

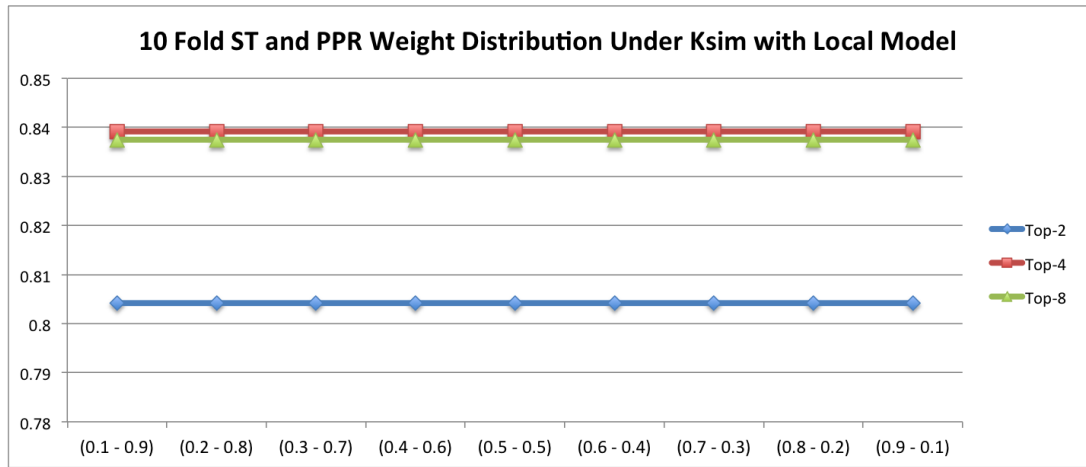


Figure B.55: 10-Fold Validation ST and PPR Weight Effects on HPR under Ksim with Local Model

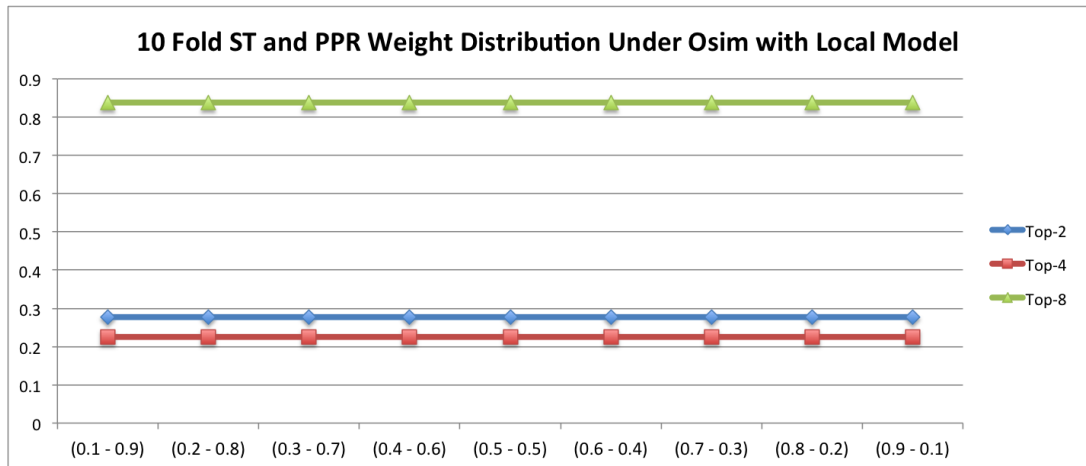


Figure B.56: 10-Fold Validation ST and PPR Weight Effects on HPR under Osim with Local Model