لقلب الدافئ إلى من نوّر لي دربي إلى نبع الحنان الذي لاينضب إلى الحضن الذي ألجأ إليه عندما أتعب إلى لمسة الأمل إلى من أرى ملامح الأصالة فيها .....

إلى أمي الحبيبة

إلى من يعمل على غرس بذور الخير في نفسي وينير نبراس الأمل في روحي إلى من يسعى لأرتاح ويدفعني نحو النجاح إلى رمز التضحية .....

إلى أبي الغالي

أسامة سليق

إلى من وضعاني على درب النجاح وهيئا لي جميع سبله .. إلى معلمي الأوليين

إلى أبي وأمي ..

إلى من ساندني وكان عونا لي..

إلى أخي ..

مررت معكم بأيام ستظل عيناي تفيض بالدمع كلما ذكرتها ..

إلى أصدقائي ..

آنا عجميان

إلى من علمني الكلمة الأولى

إلى الذي تعب وكد من اجل مستقبلي

إلى من كان مظلتي من مخاطر حياتي

إلى والدي الغالي...

إلى القلب الحنون و عشقي الأول

إلى من سهرت ليالي لتضيء دربي بالنجاح

إلى أمي الغالية...

إلى من هم أقرب إلى نفسي مني

إلى أصدقاء ماضيي وحاضري ومستقبلي

إلى أخوتي وأخواتي...

إلى من ملؤوا حياتي فرحا وسعادة

إلى من شاركوني همومي

إلى أصدقائي...

بلال الهلال الشريفي

إلى كل من كان بجانبي طوال تلك السنين ..

إلى من سكنت صورهم قلبي ..

إلى من نكن لهم خالص المحبة والاحترام ..

إلى من أدين لهم بالجميل والعرفان ..

إلى بلدي الحياة والصمود ..

إلى سوريا وأهلها ..

إلى أمي ..

إلى أبي ..

إلى أخوتي ..

لما موازيني

فهرس المحتويات

Contents

الفصل الأول مدخل إلى المشروع 23

1 مقدمة 23

2 عرض المشكلة 25

3 أهداف المشروع 26

4 إدارة المشروع 27

4.1 الموارد المتاحة 27

4.1.1 الموارد العتادية 27

4.1.2 الموارد البرمجية 27

4.1.3 الموارد البشرية 28

4.2 المهام الرئيسية 28

4.2.1 مرحلة الدراسة النظرية 28

4.2.2 مرحلة التحليل 29

4.2.3 مرحلة التصميم 29

4.2.4 مرحلة التحقيق 29

4.2.5 مرحلة الاختبارات 29

4.2.6 الجدول الزمني للمهام 30

4.3 مخططات غانت 32

4.3.1 مخطط غانت لمرحلة الدراسة النظرية 32

4.3.2 مخطط غانت لمرحلة التحليل 32

4.3.3 مخطط غانت لمرحلة التصميم 33

4.3.4 مخطط غانت لمرحلة التحقيق 33

4.3.5 مخطط غانت لمرحلة الاختبار 34

الباب الأول الدراسة المرجعية 35

الفصل الثاني مشاريع في فلترة صفحات الإنترنت 37

1 مشاريع في فلترة صفحات الإنترنت 39

1.1 DansGuardian 39

1.2 K9 42

1.3 OpenDNS 45

1.4 We-Blocker 47

1.5 Anti-Porn 48

1.6 Squid 50

1.7 SquidGuar 51

1.8 SafeSquid 53

1.9 QuintoLabs 55

الفصل الثالث دراسة مرجعية عن فلترة وتصنيف البريد الإلكتروني 59

1 إدارة Outlook للبريد الإلكتروني من خلال القواعد 60

1.10 استخدام Outlook rule 61

2 Zimbra Anti-spam system 70

الباب الثاني الدراسة النظري 73

الفصل الرابع wordnet && wordnet domains 75

1 معجم wordnet3 75

2 مجالات Extended Wordnet Domains 80

الفصل الخامس RDF && DBpedia && Linked Data 81

1 ثلاثيات RDF 81

2 DBpedia 83

3 Linked Data 84

الفصل السادس تحليل وفهم المحتوى 85

1 استبدال الضمائر بالكلمات الأصلية التي تعود عليها هذه الضمائر Coreferencing 86

2 إعادة الكلمات إلى أصلها Lemmatization 87

3 إزالة غموض الكلمات Word Sense Disambiguation 87

3.1 خورازمية Lesk التقليدية 89

3.2 التحسينات على خورازمية Lesk 91

3.2.1 استخدام شبكة دلالية 91

3.2.2 ايجاد التقاطعات بين تعريفين 94

3.2.3 استراتيجية عمل خوارزمية Global Disambiguation 95

3.3 خوارزمية Hood 99

3.3.1 فك غموض الكلمة وتصنيف المستندات 101

3.3.2 إنشاء Hood (Hood Construction) 101

3.3.3 فك الغموض 103

3.4 إزالة الغموض بالاستعانة ب WordNet Tagged glosses 103

3.4.1 لمحة عن مشروع WordNet Semantically Tagged glosses 103

3.5 الخوارزمية المتبعة لإزالة الغموض 104

3.5.1 ايجابيات و سلبيات هذه الخوارزمية 106

4 مقارنة بين الخوارزميات الأربعة المستخدمة في إزالة الغموض 106

4.1 طريقة تحليل وفهم المحتوى اعتمادا على معالجة اللغات الطبيعية 107

4.1.1 قراءة صفحة الإنترنت المطلوبة وتحويلها إلى ملف نصي 107

4.1.2 تجزئة الكلمات 107

4.1.3 إزالة المحارف الغريبة 107

4.1.4 تقسيم النص إلى جمل 107

4.1.5 المعالجة الدلالية وعملية الفلترة 107

4.2 طريقة فهم المحتوى اعتمادا على ثلاثيات RDF 109

الباب الثالث مرحلة التحليل 111

الفصل السابع وثيقة دراسة متطلبات النظام 113

1 تعريف المشكلة المدروسة 113

2 الغاية من بناء هذا المنتج 114

3 تحديد الجهة المهتمة بالمشروع 114

4 أجزاء المشروع الأساسية 114

الفصل الثامن حالات استخدام النظام 117

1 مقدمة 117

2 مخطط حالات الاستخدام الأولي 117

2.1 مخطط حالات استخدام مخدم البروكسي Smaz Proxy 118

2.2 مخطط حالات استخدام مصنّف الأخبار Smaz Filter 119

3 حالات استخدام مخدم البروكسي Smaz Proxy 120

3.1 وضع إعدادات النظام 120

3.2 تعديل إعدادات النظام 121

3.3 إضافة قاعدة فلترة 122

3.4 تعديل قاعدة فلترة 123

3.5 طلب صفحة ويب 124

3.6 حذف قاعدة فلترة 125

3.7 عرض محتوى ملف القواعد 126

3.8 عرض محتوى ملف الإعدادات 127

4 حالات استخدام إضافة المتصفح Smaz Filter 128

5 حالات استخدام مصنّف الأخبار Smaz Reader 128

5.1 عرض الأخبار التابعة لمجال معين 128

5.2 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها 129

5.3 وضع إعدادات التطبيق 130

5.4 عرض المجالات المتواجدة 131

5.5 إضافة قناة أخبار 132

5.6 عرض القنوات المتواجدة 133

5.7 حذف قناة 134

5.8 وضع إعدادات النظام 135

الفصل التاسع مخطط الصفوف المفاهيمي 139

1 مخطط الصفوف المفاهيمي لمخدم البروكسي Smaz Proxy 140

2 توصيف الصفوف المفاهيمية لمخدم البروكسي 141

3 مخطط الصفوف المفاهيمي للقارئ Smaz Reader 145

4 توصيف الصفوف المفاهيمية للقارئ 146

الباب الرابع الدراسة التصميمية 147

الفصل العاشر مخطط الصفوف 149

1 مخطط صفوف مخدم البروكسي Smaz Proxy 149

2 شرح بعض صفوف مخدم البروكسي Smaz Proxy 150

150..... hoodDisambiguator 2.1
151..... Word 2.2
151..... Combination 2.3
151..... Instance 2.4
152.....DbPediaHelper 2.5
152..... ExDomainsHelper 2.6
153..... FileReaderHelper 2.7
153.....GeneralHelper 2.8
154.....StanfordCoreNLPHelper 2.9
154.....DocDomainsFinder 2.10
155..... SynsetDomain 2.11
155..... JWIHelper 2.12
155.....SemanticAnalyzer 2.13
155..... SemanticAnalyzerImpl 2.14
156..... Smaz Reader مخطط صفوف القارئ 3
157.....Smaz Reader شرح بعض صفوف القارئ 4
157.....RssChannel 4.1
157.....RssItem 4.2
158..... Domain 4.3
158..... DataBaseHelper 4.4
159.....SemanticAnalyzerDomain 4.5
159.....WebHelper 4.6
160.....Config 4.7

الفصل الحادي عشر تصميم قاعدة المعطيات 163

1 جداول توصيف الكيانات 163

2 العلاقات بين الكيانات 165

3 مخطط قاعدة البيانات 166

الباب الخامس مرحلة التحقيق 167

الفصل اللثاني عشر تحقيق النظام 169

1 خدمة الويب Smaz web service 171

1.1 فكرة RMI 171

1.1.1 RMI Server 173

1.1.2 RMI Client 173

2 بنية ملف الإعدادات Config File 174

3 بنية ملف القواعد Rules File 177

4 مخدم البروكسي Smaz Proxy 179

4.1 Icap server 179

4.2 مخطط العمل وآلية الاتصال بين مكونات النظام 180

5 إضافة المتصفح Smaz Filter 181

6 القارئ Smaz Reader 182

اختبار أداء Smaz Proxy 182

7 البيئة البرمجية 184

7.1 لغة البرمجة المستخدمة 184
7.2 jsp web service 186
7.3 ICAP 186
7.4 المكتبات المستخدمة 186
7.4.1 Apache Jena 186
7.4.2 Apache PDFBox 187
7.4.3 Apache POI 187
7.4.4 Boilerpipe 187
7.4.5 Jsoup 188
7.4.6 JWI 188
7.4.7 Stanford CoreNLP 190
7.5 SpotLight 191
الفصل الثالث عشر واجهات النظام 193
1 واجهات مخدم البروكسي Smaz Proxy 194
1.1 مخدم icap والبرمجة بداخله 194
1.2 التأكد من أن المخدم متصل وقادر على معالجة الطلبات 195
1.3 طلب صفحة ومعالجتها وحجبها 196
وضع إعدادات المخدم 197
2 إضافة المتصفح Smaz Filtering 198
2.1 واجهة الإضافة التي يتم من خلالها وضع الإعدادات 198

2.2 معلومات عن الإضافة 199

2.3 إضافة Smaz Filter إلى قائمة إضافات المتصفح 199

3 تطبيق الموبايل Smaz Reader 200

3.1 واجهة عرض المجالات المختارة مسبقا 200

3.2 واجهة عرض الأخبار الخاصة بمجال معين 201

3.3 واجهة إعدادات التطبيق 202

3.4 واجهة عرض قنوات الأخبار 203

3.5 واجهة تعديل قناة أخبار 204

3.6 واجهة توضح اختيار المجالات المطلوب عرضها 205

3.7 واجهة توضع معلومات عن التطبيق 206

4 محرر القواعد Rule editor 207

4.1 واجهة إضافة قاعدة 207

4.2 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعة 208

4.3 واجهة عرض القواعد الموجودة والتعديل عليها 209

الباب السادس مرحلة الاختبارات 211

1 مقدمة 213

1الاختبارات عبر مراحل تطوير النظام 213

1.1 اختبار الواحدة البرمجية 213

1.1.1 اختبار التكامل 213

1.1.2 اختبار النظام 214

1.1.3 اختبارات الجودة 214

الباب السابع الآفاق المستقبلية، الملحقات والمراجع 217

الفصل الرابع عشر الآفاق المستقبلية 219

الفصل الخامس عشر الملحق أ– مسرد المصطلحات 221

aالفصل السادس عشر الملحق ب– فهرس المخططات 225

الفصل السابع عشر الملحق ج– فهرس الجداول 229

الفصل الثامن عشر الملحق د– المراجع 231

تمهيد

تشهد الشبكة العنكبوتية اليوم تطوراً ملحوظاً وتزايداً في أعداد المواقع التي تنضم يوميا إلى لائحة المواقع على شبكة الإنترنت، والتي تتنوع في محتواها فمنها الترفيهي ومنها الثقافي ومنها التعليمي فضلا عن شبكات التواصل الإجتماعي وغير ذلك من المجالات، هذا النمو المتزايد جعل من الصعب السيطرة على محتوى صفحات الإنترنت وضمان أن ما يتم تصفحه هو ضمن مجال معين نرغب به.

من هنا أتت الحاجة إلى فلترة صفحات الإنترنت، فمثلا لضمان عدم القدرة على الوصول إلى المواقع الإباحية يتم حظر قائمة من المواقع الإباحية وبالتالي عند طلب المستخدم لهذا الموقع لن يتم عرض تلك الصفحة من قبل المتصفح، لكن هذه الطريقة في الحجب غير مجدية نظراً لعدم قدرتنا على حصر المواقع التي تندرج تحت مجال معين، فضلا عن ظهور مواقع جديدة يومياً، أو أن يتم الحجب من خلال تحديد قائمة من الكلمات الغير مرغوب بها وبالتالي عند طلب صفحة تحوي كلمة من هذه الكلمات لن يتم عرض محتواها، لكن هذه الطريقة ليست عملية أيضا ففي حالات قد ترد كلمات مرادفة للكلمات الغير مرغوب بها ومع ذلك سيتم عرض الصفحة لعدم ورود نفس الكلمات التي تم تحديدها سابقا، وبالتالي نحن بحاجة إلى طريقة فلترة آنية تتم لحظة طلب الموقع لتتم عليه عمليات معالجة وبناءً على تلك المعالجة الدلالية يتم عرض الصحة أم لا.

نقدم في هذا المشروع توصيفا لطريقة عملية آنية تعتمد على الفهم الدقيق لمحتوى صفحة الإنترنت لحظة طلبها من قبل المستخدم حيث نقوم بالحصول على النص الموجود ضمن صفحة الإنترنت وفهمه من خلال عمليات معالجة نحوية دلالية توصلنا في النهاية إلى فهم محتوى الصفحة ومعرفة المجالات التي تدور حولها وفق نسب مئوية تعبر عن مدى انتماء الصفحة إلى كل مجال وبناءً على تلك النسب يتم أخذ قرار عرض تلك الصفحة للمستخدم أم لا وفق قواعد مسبقة موضوعة من قبل خبراء وبما يتماشى مع حاجات الشركة المستخدمة لهذا النظام.

الفصل الأول

مدخل إلى المشروع

1 مقدمة

بدأت المكتبات باستخدام فلاتر الإنترنت في أواخر 1990 بسبب ضغوط المجتمعات وقانون حماية الإنترنت للطفولة Children's Internet Protection Act (CIPA) وهو القانون الاتحادي الذي يتطلب من جميع أجهزة الكمبيوتر في المكتبة العامة أن تحوي فلاتر لصفحات الإنترنت التي سيتم طلبها من قبل مستخدمي هذه المكانب، حيث تقول إحدى الدراسات الإحصائية أن نسبة استخدام الفلاتر في المكتبات قد ازدادت من 25٪ في عام 2000 إلى 65٪ في عام 2005.

في عام 1996 ، أصدر الكونغرس الأميركي قانون حظر المواقع الإباحية والمواقع ذات المحتوى السيء من الإنترنت، وفي عام 1997 وقفت جماعة الحريات المدينة ضد ذلك القانون وألغي بعد ذلك، وهكذا أصبح من الضروري تصفية المحتوى.

منذ s1990، وقفت العديد من المجموعات ضد استخدام الشركات لتصفية المحتوى، نددت بالاستخدام المفرط لفلاتر الإنترنت وبالتالي حجب مواقع ليس هناك أسباب واضحة وراء حجبها، ولكن هناك حاجة ضرورية لفلترة المحتوى من قبل الآباء لحماية أطفالهم من المواقع المسيئة، أصبحت شركات تجارية كبرى تعمل على فلترة المحتوى، وقد شهد السوق نموا مضطردا ضمن هذا المجال، وحتى نمو 20 في المئة خلال اقتصاد الركود في عام 2009.

تطورت خلال هذه الفترة طرق الفلترة المستخدمة بدءا من الحجب اعتمادا على قائمة سوداء تضم مجموعة من المواقع التي لا نرغب بعرضها، ومن ثم الحجب من خلال مجموعة من الكلمات لمجرد ورود إحدى هذه الكلمات ضمن صفحة الإنترنت المطلوبة لن يتم عرض الصفحة، والحجب اعتمادا على meta tags محددة لمجرد ورود هذه الكلمات المفتاحية ضمن الكلمات المفتاحية الخاصة بالصفحة يتم الحجب، والحجب اعتمادا على تعابير نظامية وغير ذلك من الطرق المستخدمة في الفلترة.

تعاني هذه الطرق من عدم جديتها ودقتها فلن نستطيع بهذه الطرق الحجب الفعلي لمحتوى كامل غير مرغوب به وبالتالي لابد من طرق أكثر عملية وفعالية بنفس الوقت، هنا لابد من الفهم الدلالي الدقيق لمحتوى الصفحة واتخاذ قرار الحجب لحظة طلب الصفحة بناءً على محتواها.

2 عرض المشكلة

بعد التزايد الكبير لمواقع الإنترنت وعدم القدرة على إحصائها، أصبح من غير الممكن متابعة كل مواقع الإنترنت ومعرفة محتواها ومضمونها، وبالتالي أصبح من الصعب القيام بعمليات مسبقة لمعالجة المحتوى من فلترة وتصنيف وغير ذلك من العمليات.

فلترة صفحات الإنترنت كانت عبارة عن حصر المواقع التي نريد حجبها ووضعها ضمن قائمة سوداء وعند طلب صفحة موجودة ضمن القائمة يتم حجبها عن المستخدم، لكن هذه الطريقة لم تعد مجدية كون الشبكة العنكبوتية في تطور ونمو مستمر لايمكن السيطرة عليه.

فلترة صفحات الإنترنت من خلال قائمة من الكلمات الغير مرغوب بها غير مضمونة أيضا، فيمكن ورود كلمات ضمن صفحة الإنترنت المطلوبة بمعنى الكلمة الغير مرغوب بها ويتم عرض الصفحة لعدم ذكر الكلمة صراحة.

إذا كل عمليات معالجة المحتوى من قلترة وتصنيف وغير ذلك تتطلب فهم للمحتوى، عملية الفهم المسبق لكل المحتوى عملية ليست منطقية ولا يمكن تطبيقها، واللجوء إلى عمليات المعالجة اليدوية غير مجدي.

3 أهداف المشروع

يهدف المشروع إلى تقديم خدمة تعمل على فهم المحتوى بشكل آلي آني، تقوم هذه الخدمة باستخراج النص من الصفحة المطلوبة، معالجة هذا النص من خلال إجراء عمليات معالجة لغات طبيعية متتالية، ومن ثم الحصول على معاني كلمات من خلال الاستعانة بالمعاجم والأنطولوجيات المشهورة، ومن ثم الحصول على المجالات التي تنتمي لها الصفحة، أي معرفة أهم المواضيع والمجالات التي تتحدث عنها الصفحة المطلوبة، بهذه الطريقة بتنا قادرين على الفهم الدلالي لمحتوى الصفحة ومعرفة مضمونها، هذا الفهم الدلالي يتم بشكل آلي وبالتالي تم التخلص من طرق المعالجة اليدوية الغير مجدية، بهذه الطريقة بتنا قادرين على القيام بعمليات على المحتوى من فلترة وتصنيف وغير ذلك.

يقدم المشروع إذاً طريقة دلالية آلية آنية لفهم المحتوى، ومن ثم تم الاستفادة من هذه الطريقة في فلترة المحتوى من خلال تطبيقين proxy server و chrome extention ، وتصنيف المحتوى من خلال تطبيق موبايل يتم فيه تصنيف الأخبار ضمن مجموعة من المجالات كما سنرى لاحقا.

4 إدارة المشروع

نعرض في هذا الفصل الموارد المتنوعة البشرية والعتادية التي احتاجها المشروع، ثم سنقوم بعرض المهام الرئيسية في المشروع، إضافة إلى تحديد الفترة اللازمة لتحقيق كل مهمة وذلك من خلال جدول يوضح الفترة الزمنية المحددة لكل مهمة.

4.1 الموارد المتاحة

4.1.1 الموارد العتادية

أربعة حواسيب محمولة مجهزة بنظام Microsoft windows 7 وحاسوبان منهما مجهزان بنظام التشغيل Linux.

4.1.2 الموارد البرمجية

- بيئة netbeans java لتطوير البرمجيات؛
- بيئة Eclipse؛
- Java for android؛
- بيئة java script؛
- برنامج Enterprise Architect لتصميم النظام هندسيا.

4.1.3 الموارد البشرية

- فريق عمل المشروع

تضمن فريق العمل أربع أشخاص:

  - أسامة سليق؛
  - آنا عجميان؛
  - بلال الهلال الشريفي؛
  - لما موازيني.

- الإشراف العلمي

تضمن فريق الإشراف العلمي:

  - الدكتور خليل عجمي؛
  - الدكتور سامي خيمي؛
  - المهندس محمد السّاطي.

4.2 المهام الرئيسية

4.2.1 مرحلة الدراسة النظرية

- دراسة الخوارزميات المستخدمة في إزالة غموض الكلمات؛
- دراسة حول طرق فهم النص دلاليا؛
- دراسة عن أشهر البرامج المستخدمة في فلترة صفحات الإنترنت وطرق عملها.

4.2.2 مرحلة التحليل

- وضع حدود المشروع؛
- تحديد متطلبات المشروع؛
- تحديد حالات الاستخدام.

4.2.3 مرحلة التصميم

- تصميم قاعدة معطيات النظام؛
- تصميم الصفوف اللازمة لتحقيق النظام.

4.2.4 مرحلة التحقيق

- تحقيق قاعدة المعطيات؛
- تحقيق صفوف النظام؛
- تحقيق الرماز المصدري.

4.2.5 مرحلة الاختبارات

- اختبار الوحدات البرمجية؛
- اختبار التكامل؛
- اختبار النظام؛
- اختبار الجودة .

4.2.6 الجدول الزمني للمهام

نعرض فيما يلي جدول زمني يوضح المدة الزمنية لكل مرحلة من مراحل المشروع حيث تتضمن كل مرحلة التقرير والتوثيق المرافق للعمل.

الرقم	المهمة	المدة	تاريخ البدء	تاريخ الانتهاء
1	دراسة خوارزميات إزالة غموض الكلمات.	15	1/3/2013	15/3/2013
2	دراسة حول فهم النص دلاليا.	20	16/3/2013	5/4/2013
	دراسة عن برامج فلترة الإنترنت وطرق عملها.	10	6/4/2013	15/4/2013
3	وضع حدود المشروع.	5	16/4/2013	20/4/2013
4	تحديد متطلبات المشروع.	10	21/4/2013	30/4/2013
5	تحديد حالات الاستخدام.	5	1/5/2013	5/5/2013
6	تصميم قاعدة معطيات النظام.	5	6/5/2013	10/5/2013
7	تصميم الصفوف اللازمة لتحقيق النظام.	10	11/5/2013	20/5/2013
8	تحقيق قاعدة المعطيات.	5	21/5/2013	25/5/2013
9	تحقيق صفوف النظام.	5	26/5/2013	30/5/2013
10	تحقيق الرماز المصدري	20	1/6/2013	20/6/2013

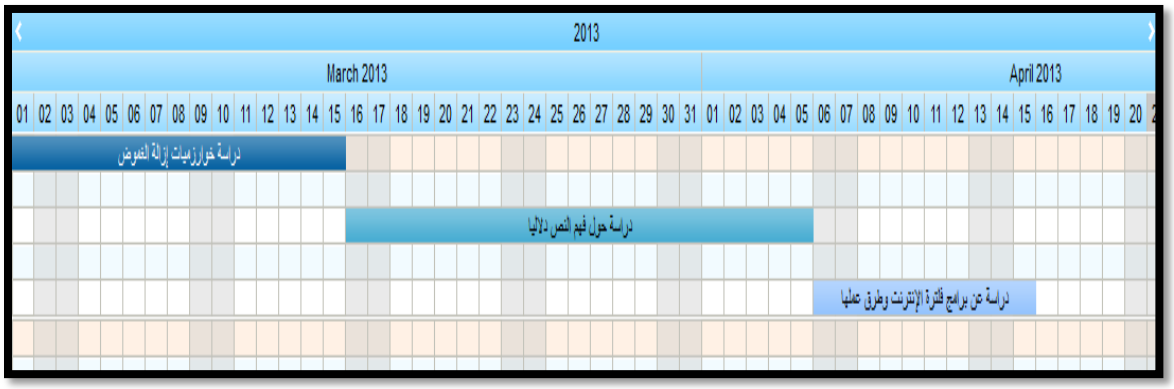
11	مرحلة الاختبارات	10	21/6/2013	30/6/2013
12	العرض التقديمي	10	1/7/2013	10/7/2013

جدول 1 توزيع المهام خلال فترات زمنية محددة

4.3 مخططات غانت

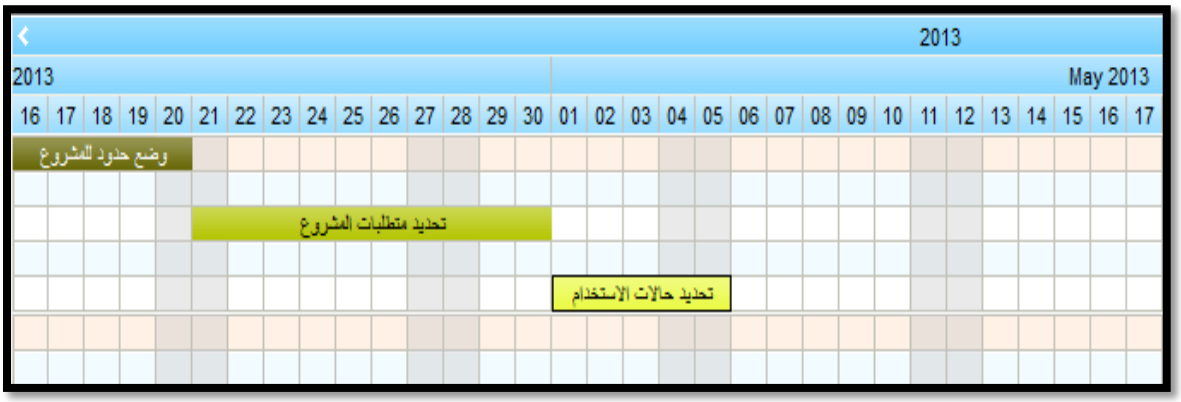
تظهر هذه المخططات توزع المهام زمنيا والعلاقات المتبادلة بين هذه المهام.

4.3.1 مخطط غانت لمرحلة الدراسة النظرية



الشكل 1 مخطط غانت لمرحلة الدراسة النظرية

4.3.2 مخطط غانت لمرحلة التحليل

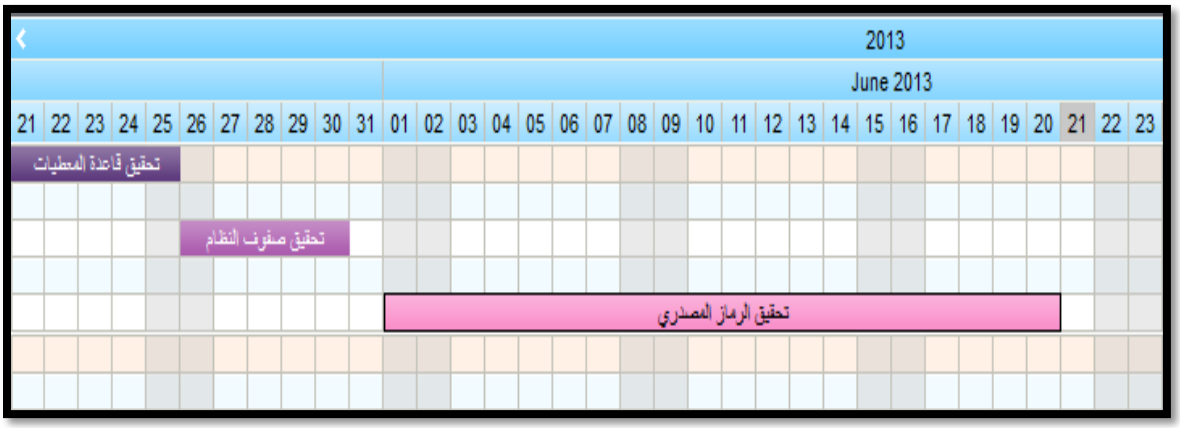


الشكل 2 مخطط غانت لمرحلة التحليل

4.3.3 مخطط غانت لمرحلة التصميم

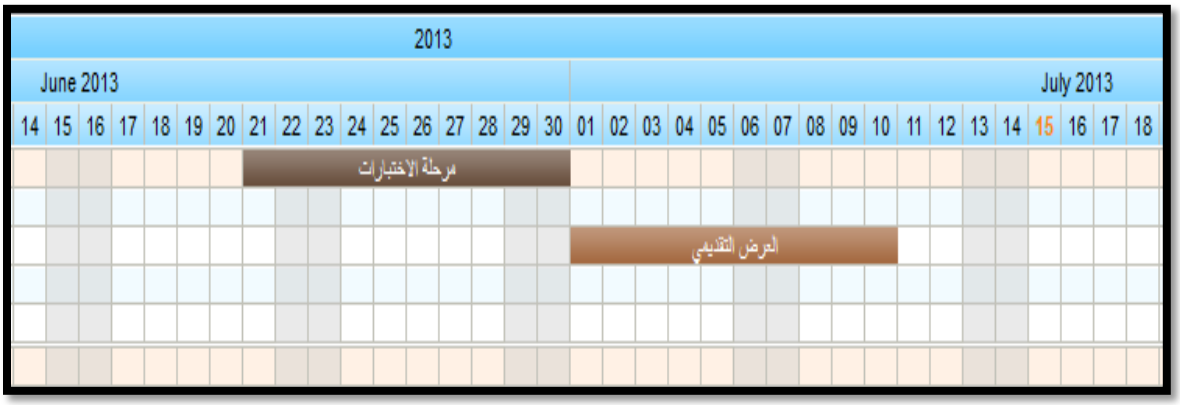
الشكل 3 مخطط غانت لمرحلة التصميم

4.3.4 مخطط غانت لمرحلة التحقيق



الشكل 4 مخطط غانت لمرحلة التحقيق

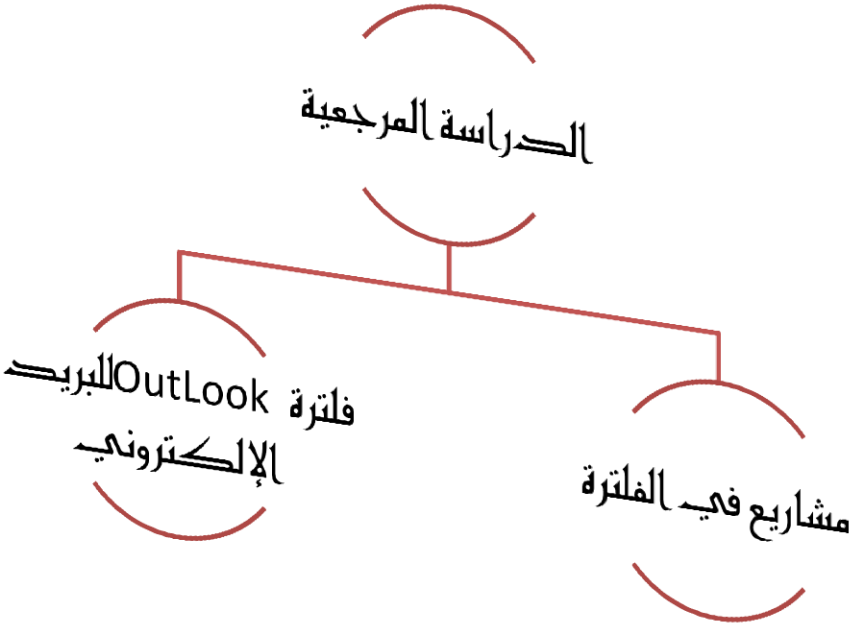
4.3.5 مخطط غانت لمرحلة الاختبار



الشكل 5 مخطط غانت لمرحلة الاختبار

الباب الأول

الدراسة المرجعية

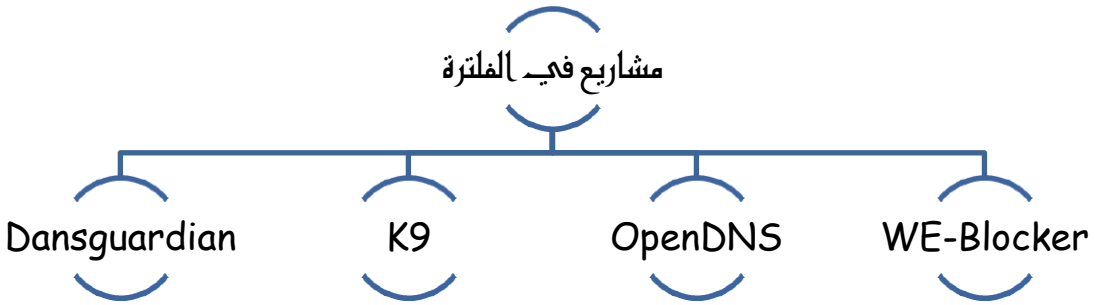


الشكل 6 النقاط الأساسية في الدراسة المرجعية

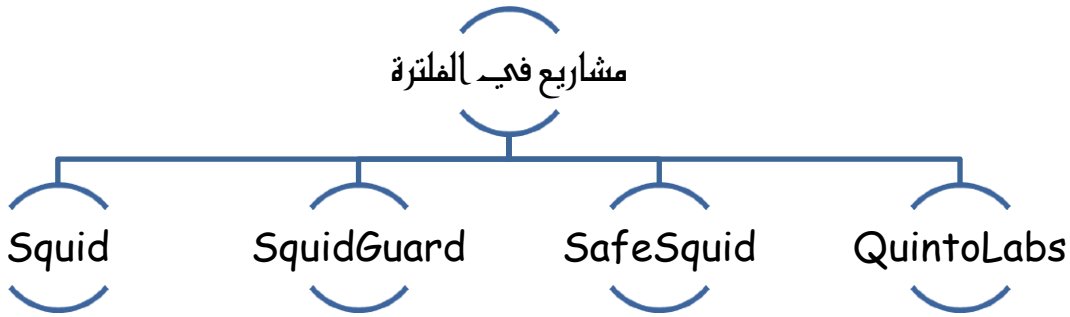
الفصل الثاني

مشاريع في فلترة صفحات الإنترنت

يتناول المشروع مجموعة من التطبيقات الناتجة عن فهم وتحليل للمحتوى تم استخدامه في فلترة صفحات الإنترنت من خلال تطبيقين اثنين، كما تم استخدامه في تصنيف المحتوى من خلال تطبيق موبايل، نعرض في هذا الفصل مجموعة من البرامج والتطبيقات التقليدية المستخدمة في فلترة صفحات الإنترنت، لنرى كيف تتم الفلترة في الواقع وفي معظم الأحيان نجد أن الطرق المستخدمة هي طرق تقليدية، وأن الخوارزميات المتبعة هي خوارزميات بسيطة لا تفي بالغرض المرجو.



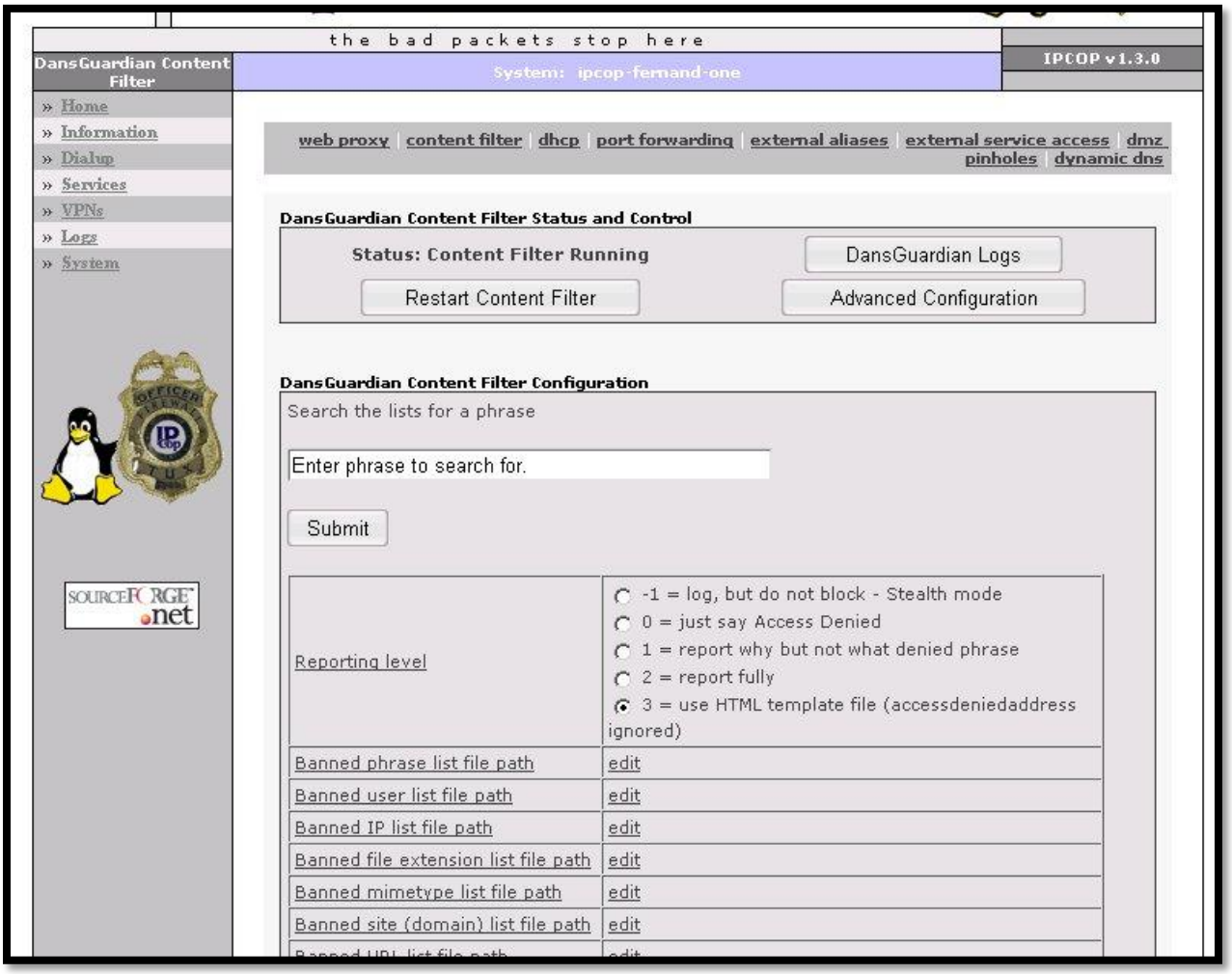
الشكل 7 النقاط الأساسية في مشاريع الفلترة



الشكل 8 النقاط الأساسية في مشاريع الفلترة

1 مشاريع في فلترة صفحات الإنترنت

1.1 (Cross Platform, Free) DansGuardian



الشكل 9 صورة توضح واجهة DansGuardian

- مجاني (في حال الاستخدام غير التجاري) ومفتوح المصدر؛
- يدعم أنظمة التشغيل التالية: Linux, FreeBSD, OpenBSD, NetBSD, Mac OS X, HP-UX, Solaris؛

- يعمل كإضافة على بعض مخدمات البروكسي مثل Squid؛
- يقوم بالفلترة من خلال:

1- فلترة الصفحات ليس فقط بناءً على قوائم عناوين معدة مسبقاً كما تفعل باقي الفلاتر، و إنما أيضاً يقوم بالفلترة بناءً على محتوى الصفحات؛

2- يتيح إمكانية الفلترة بناء على العناوين URL filtering باستخدام black/white lists أو باستخدام regular expressions و هنا يتم التمييز بين نوعين للفلترة:

a. فلترة موقع معين بشكل كامل؛

b. فلترة صفحات معينة من موقع ما وهذه الطريقة في الفلترة غير مجدية كثيراً بسبب التزايد المستمر و السريع لمواقع الإنترنت و صعوبة تعديل قوائم الحجب black/white lists.

3- يتيح إمكانية فلترة المحتوى باستخدام regular expressions؛

4- يتيح إمكانية فلترة المحتوى باستخدام ما يسمى Phraselists، حيث أن DansGuardian بحالته الإفتراضية يأتي مضمناً بأعداد ضخمة من هذه العبارات مقسمة إلى أصناف رئيسية، و بالطبع يمكن تعديلها، و لهذه العبارات نوعين:

a. Banned Phrases

في حال ورود أي عبارة من هذه العبارات في الصفحة يتم حجب الصفحة بشكل كامل دون أخذ أي شيء آخر بعين الاعتبار، لذلك يجب توخي الحذر عند التعديل على العبارات من هذا النوع.

مثلا في حال ورود كلمة sex يتم حجب الصفحة بشكل كامل، وتتم صياغة هذا النوع على الشكل:

<sex>

b. Weighted Phrases

هنا يتم مقابلة كل عبارة بمجموع نقط، ويتم في النهاية حساب مجموع النقط التراكمي للصفحة و في حال تجاوزت حد معين (يمكن تعديله) يتم حجب الصفحة.

مثلا إذا وردت كلمة sex مع كلمة education هذا سيخفف من الأثر السلبي وبالتالي يجب إضافة عدد سالب إلى المجموع التراكمي للصفحة و يتم صياغة العبارة من هذا النوع على الشكل التالي:

<sex><50>

<sex>,<education><-30>

أي يتم طرح 30 من المجموع التراكمي للصفحة في حال وردت كلا الكلمتين السابقتين في الصفحة وهذه الطريقة وضوحاً تعتبر أكثر دقة من سابقتها.

5- يتيح إمكانية تعديل ال URL أو محتوى الصفحات في بعض الأحيان، مثلا في حال البحث في Google يمكن أن يتم تعديل ال URL لإجبار Google Safe Search على العمل حتى لو لم يقوم الشخص الذي طلب الصفحة بتفعيل هذه الميزة صراحة؛

6- يتيح إمكانية تفعيل ميزة POST limiting

والتي تسمح بالتحكم برفع الملفات أو حجبها بشكل كامل؛

7- PICs web-site rating filtering

أي يمكن أن تتم الفلترة على مستوى الصور؛

8- Anti-virus filtering

يتم اختبار الملفات قبل تحميلها فيما إذا كانت خبيثة أم لا؛

9- Meta tags filtering

تتم الفلترة بناء على ال meta tags الخاصة بال html؛

10- File extension and file type (MIME) filtering

أي يتم حجب تحميل الملفات و الصفحات التي لها أنماط محددة يتم تحديدها مسبقاً.[1]

1.2 K9 (Windows/Mac, Free)

الشكل 10 صورة توضح واجهة K9

- مجاني من أجل الاستخدام الفردي والمنزلي، غير مجاني من أجل الشركات؛
- لا يعمل مع نظام التشغيل Linux؛
- يدعم أنظمة التشغيل Windows, Mac كبرنامج، وعلى أنظمة الموبايل مثل Ios و Android كمتصفح؛

- يقوم K9 بالفلترة من خلال:

1- Web Site Exceptions Blocks منع أو السماح بمواقع معينة تختارها؛

2- URL Keywords Blocks تحديد كلمات معينة وعند محاولة الدخول على اسم موقع يحتوى هذه الكلمة سيتم منع الدخول تلقائياً، فعلى سبيل المثال لنفرض حذفنا كلمة "casino" فعند ورود URL حاوي على هذه الكلمة مثل bestCasino.com فيتم حذفه مباشرة كونه يحتوي على كلمة محجوبة؛

3- Time Restrictions Blocks تحديد أوقات الدخول إلى الإنترنت؛

4- Timeout Blocks تحديد عدد المواقع الغير مسموح الدخول إليها خلال فترة زمنية محددة في حال تجاوز ذلك العدد يتم مثلا حجب الدخول إلى أي موقع لمدة نصف ساعة على سبيل المثال؛

5- Web Categories to Block تحديد مستوى الفلترة العام هناك خيارات:

High: يمنع تقريبا كل شيء والشات والـ yo tube

Default: الإفتراضي، يمنع المحتوى السيء، فمثلا يقوم بحجب المواقع بناءً على تصنيف مسبق بحيث يقوم بحذف موقع playboy.com كونه مصنف على أنه من النوع Pornography.

كما يمكنه منع حجب المواقع المصنفة على أنها Sex Educations مثلا من خلال تعديل بسيط في الإعدادات كالسماح بفتح كل المواقع ذات المحتوى التعليمي والشرط أن يملك الشخص الذي يقوم بالتعديل صلاحيات الإدارة؛

6- يمكن إعداد فلتر خاص بالمستخدم عن طريق Custom فمثلا يقوم المستخدم بتحديد التصنيفات التي يرغب بحجبها والمواقع التي يريد والكلمات التي يريد حجبها من عنوان المواقع؛

7- عند طلب URL لموقع ما غير محجوب صراحة ولا يحوي على كلمات محجوبة:

- ✓ يقوم K9 بالبحث عن الموقع المطلوب ضمن الذاكرة الخاصة به cache memory، في حال حدوث مطابقة لا يقوم بسؤال BCWF عن تصنيفه؛
- ✓ وإلا يقوم بطلب استعلام من BCWF لتقوم بدورها بالبحث عنها ضمن قاعدة المعطيات الخاصة بها، في حال حدوث مطابقة تقوم بإرسال التصنيف إلى K9، وإلا سيقوم BCWF بجلب محتوى الصفحة وعمل تقييم آني لها بناء على DRTR، حيث يقوم بعملية تصنيف مباشرة على الموقع عن طريق خوارزميات تحليلية سريعة اعتماداً على خوارزمية ذكية عالية المستوى هي عملية التعليم Machine Learning، حيث يعتمد على الكم الهائل من المواقع المتوفرة لديه كعينات تدريب.

وهذه العملية لا تأخذ وقت أكثر من 70 ثانية، وهي تعمل بشكل ممتاز على المواقع التي تكون من النوع Pornography حيث أن هذا النوع من المواقع هو الأكثر تقييماً ضمن قاعدة البيانات الخاصة بهذا النظام، وأيضا فهو يعتمد على links الموجودة ضمن الصفحات بحيث يبحث عن هذه الترابطات ضمن قاعدة بياناته الخاصة، وفي حال وجد أن هذه الترابطات تشير إلى أحد المواقع المخزنة لديه يقوم بمقارنته معه مباشرة ويعطيه نفس التصنيف. [2]

Dynamic Real-Time Rating™ (DRTR)

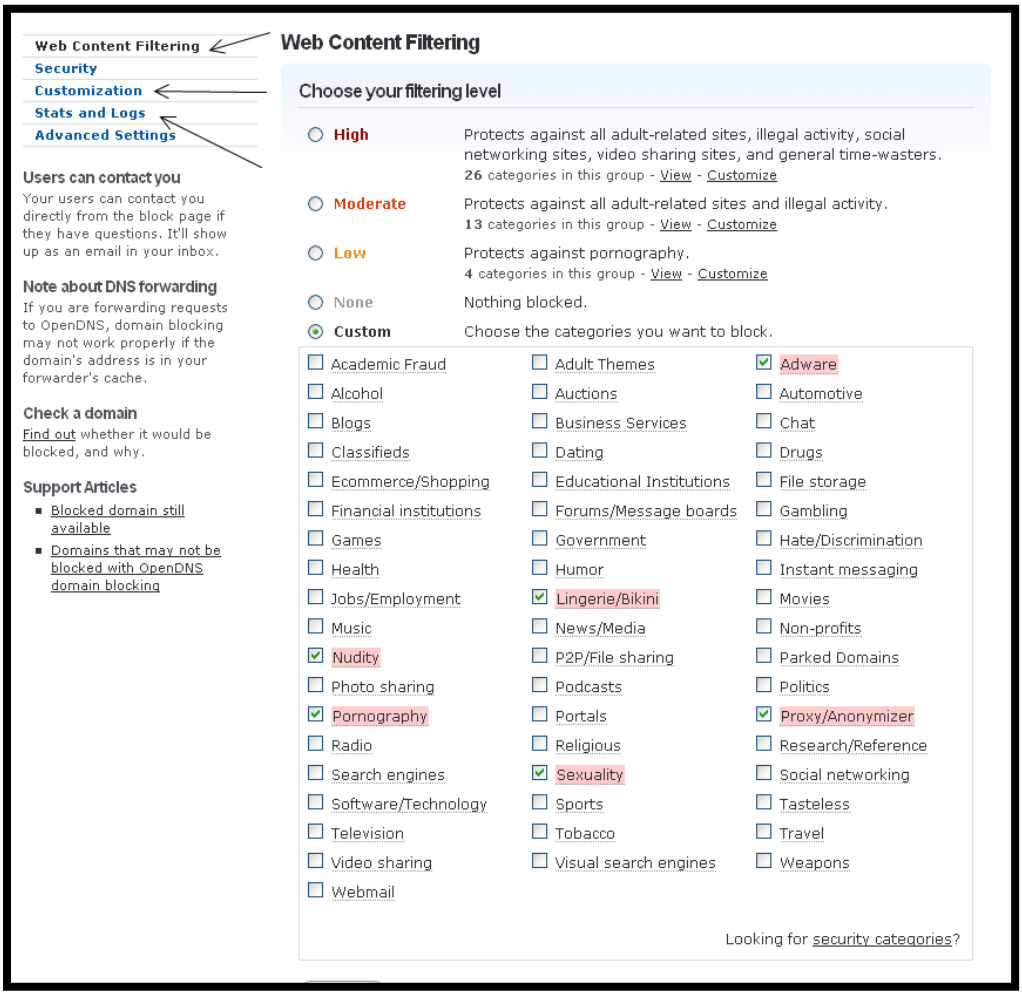
- ✓ التقييم الديناميكي في الوقت الحقيقي؛
- ✓ هي طريقة قوية ومتينة وأقوى من طرق الفلترة التقليدية المعتمدة على الكلمات المفتاحية؛
- ✓ تعتمد على خوارزمية تحليل إحصائي وذكاء اصطناعي لتقييم صفحات الإنترنت الجديدة التي لم يتم تقييمها مسبقا؛
- ✓ تستطيع ال **DRTR** تحديد تصنيف صفحة الإنترنت من دون تدخل بشري؛

1.3 (Cross Platform, Free) OpenDNS

الشكل 11 صورة توضح واجهة OpenDNS

- خدمة توفر DNS Server مجاني ولكن أيضاً بميزات تمنحها أغلب موفري الخدمات بمقابل مادي وهي "فلترة المحتوى"؛
- تقوم بكتابة ips الخاصة بال DNS Server فى Gateway سواء كان حاسوب أو راوتر، وتقوم بالتسجيل في الموقع وتدوين ip خاص بالشبكة وتعطي تلك الشبكة اسم، وبعد ذلك تحدد ما هو المرغوب به وماهو غير المرغوب، وفي حال حاول أحد الدخول على أحد الأشياء التى قمت بمنعها يتم توجيه الى صفحة تخبر برسالة معينة دون تحميل الصفحة المطلوبة؛
- القدرة على حظر المواقع عن كل المتصلين بالشبكة سواء عن طريق حاسب أو هاتف أو أي جهاز لديه القدرة على الإتصال بالإنترنت؛
- يمكن إرغام مستخدمي الشبكة عن كتابة عناوين معينة بأن توجههم إلى ips معينة مثلا إذا كتب شخص ما microsoft.con يتم توجيهه إلى kernel.org ؛
- إمكانية حجب مواقع تحتوي كلمات معينة و تجاهل الآخر؛ كما أن هناك خاصية المواقع المسموحة والممنوعة؛
- الحماية من هجمات السيرفرات و الإعلانات الضارة و من بعض الثغرات في المواقع؛

- القدرة على مراقبة الشبكة الخاصة ومعرفة أكثر المواقع تصفحاً وأكثرها حظراً وأكثر الأوقات ازدحاماً وما إلى ذلك؛
- توفر بعض التسهيلات فى عملية التصفح مثل البحث التلقائي وتصحيح بعض الأخطاء فى العناوين وإمكانية وضع الاختصارات؛
- هذا بالإضافة إلى أنها تقوم بوظيفة الـ DNS الإعتيادية مع العلم أن الـ DNS سريع ومستقر؛
- **إمكانية تحديد تصنيفات ليتم حظرها كما هو موضح في الشكل:**



الشكل 12 صورة توضح التعامل مع openDNS

- في حال كان هناك بعض المواقع الاستثنائية التي تريد منعها ولا تريد منع تصنيفاتها كاملة فيمكن عمل ذلك، كما يمكن إلغاء حظر بعض المواقع التي تم حظر تصنيفاتها كما فى الشكل:
- **إمكانية** تحديد شكل رسالة الحظر من حيث النص الذى سيظهر والصورة المعروضة؛
- قاعدة البيانات الخاصة بالخدمة لم تنشأ إلاعن طريق إبلاغ المستخدمين عن المواقع، وعند وصول من بلغوا عن الموقع بتصنيف معين إلى عدد معين يتم قبول هذا التصنيف؛
- مجاني ومفتوح المصدر.[3]

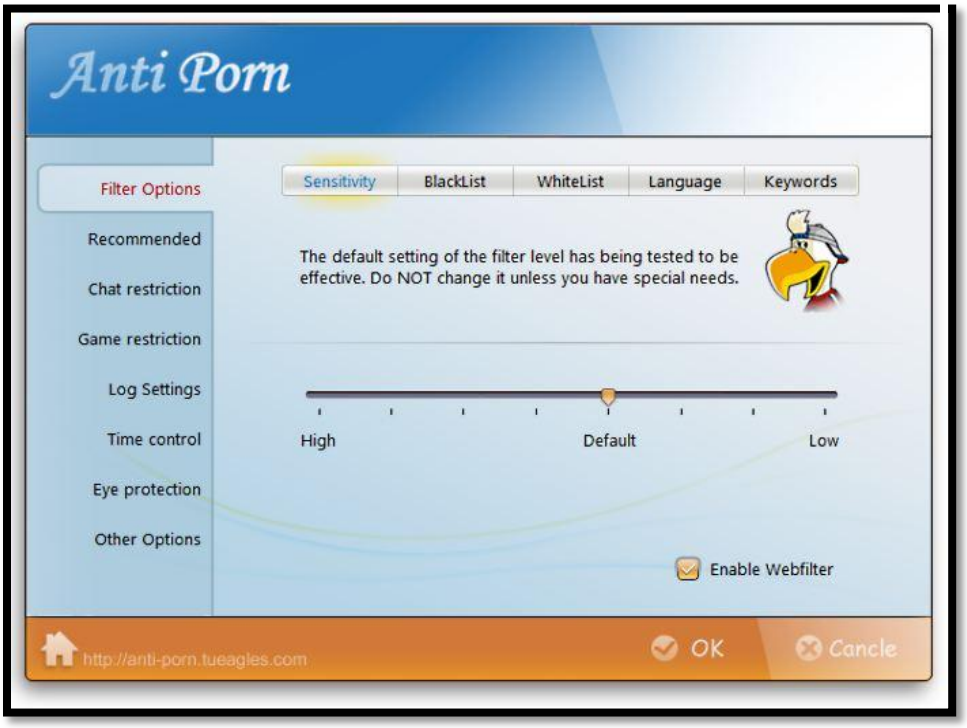
1.4 (Windows XP or earlier, Free) We-Blocker

الشكل 13 صورة توضح واجهة We-Blocker

- يتم تسجيل نشاط المستخدم على الإنترنت (المواقع التي تم دخلوها)؛

- يتم حصر دخول المستخدم إلى المواقع على الإنترنت ما عدا تلك المخزنة بقائمة المواقع المسموح بزيارتها والمسماه بالقائمة البيضاء؛
- يتم تصنيف المواقع على حسب محتوياتها وبالتالي عند وضع علامة صح أمام تصنيف منها فيتم حجب الدخول إلى المواقع التي تندرج تحت هذا التصنيف؛
- يتم حجب المواقع التي تحتوي على كلمات سيئة غير مرغوب فيها؛
- الإقلال من وضع حدود على إستخدام الإنترنت لأن ذلك سيقلل من سرعة تحميل الصفحات بالإضافة إلى صعوبة الدخول إلى المواقع في بعض الأحيان.

1.5 Anti-Porn (Windows 2000/2003/XP/Vista/7)



الشكل 14 صورة توضح واجهة anti-porn

- حجب المواقع الغير مرغوب فيها وبرامج الشات؛
- الإغلاق قبل العرض، يقوم Anti-Porn بالإطلاع السريع على كل محتويات الصفحة قبل فتحها ويغلق الصفحات الممنوعة؛
- إمكانية إضافة أي موقع إلى القائمة السوداء ليتم حجبها؛
- إمكانية استثناء أي موقع قام البرنامج بحجبه من خلال إضافة رابط الموقع إلى القائمة البيضاء؛
- إمكانية تحديد اللغات التي سيتعامل معها البرنامج للتحكم في فلترتها؛
- إمكانية إضافة أي كلمة إلى قائمة الكلمات المفتاحية ليقوم البرنامج بحجب المواقع المتضمنة لهذه الكلمة تلقائيا؛
- تحديد الدردشة على الإنترنت، يمكن لـ Anti-Porn تحديد استخدام برامج شائعة للدردشة كما يمكنه تحديد الدخول الى غرف دردشة أساسية؛
- تحديد أوقات الدخول إلى الإنترنت باليوم والساعة؛
- سجل التاريخ، يحتفظ Anti-Porn بسجل كامل لكل المواقع التي تم زيارتها سواء تم تنقيتها أم لا؛
- ينصح بالمواقع الموثوق فيها، Anti-Porn يقوم بترشيح عدد من المواقع المناسبة مع إمكانية إضافة المزيد من المواقع التي تريد؛
- مجاني ومفتوح المصدر.

1.6 Squid

الشكل 15 صورة توضح واجهة squid

- خادم بروكسي يعمل مع أنظمة تشغل Linux و Unix وحاليا قامت شركة بإعداده للعمل ضمن الويندوز؛
- مستخدم من قبل مئات من مزودي خدمة الإنترنت لتخديم مستخدميه بأفضل الطرق للوصول إلى الإنترنت؛
- يقوم بمراقبة المعطيات المنقولة بين المخدم والزبون والعمل على تحسين الأداء وتخزين الصفحات التي تم زيارتها مؤخراً والأكثر زيارة مع تحديد المستخدم والوقت والتاريخ بدقة؛
- مسرع إنترنت باستخدام خاصية التخزين المؤقت caching وفقاً للبروتوكولات HTTP, HTTPS, FTP, Gopher؛
- يقوم بتوزيع الإنترنت على الأجهزة التابعة للشبكة حسب الإعدادات الخاصة فيه؛

- حجب المواقع التي لا تريد أن تتصفحها أجهزة الشبكة؛
- فلترة المواقع والكلمات وتحديد سماحيات الوصول؛
- مجاني ومفتوح المصدر.[4]

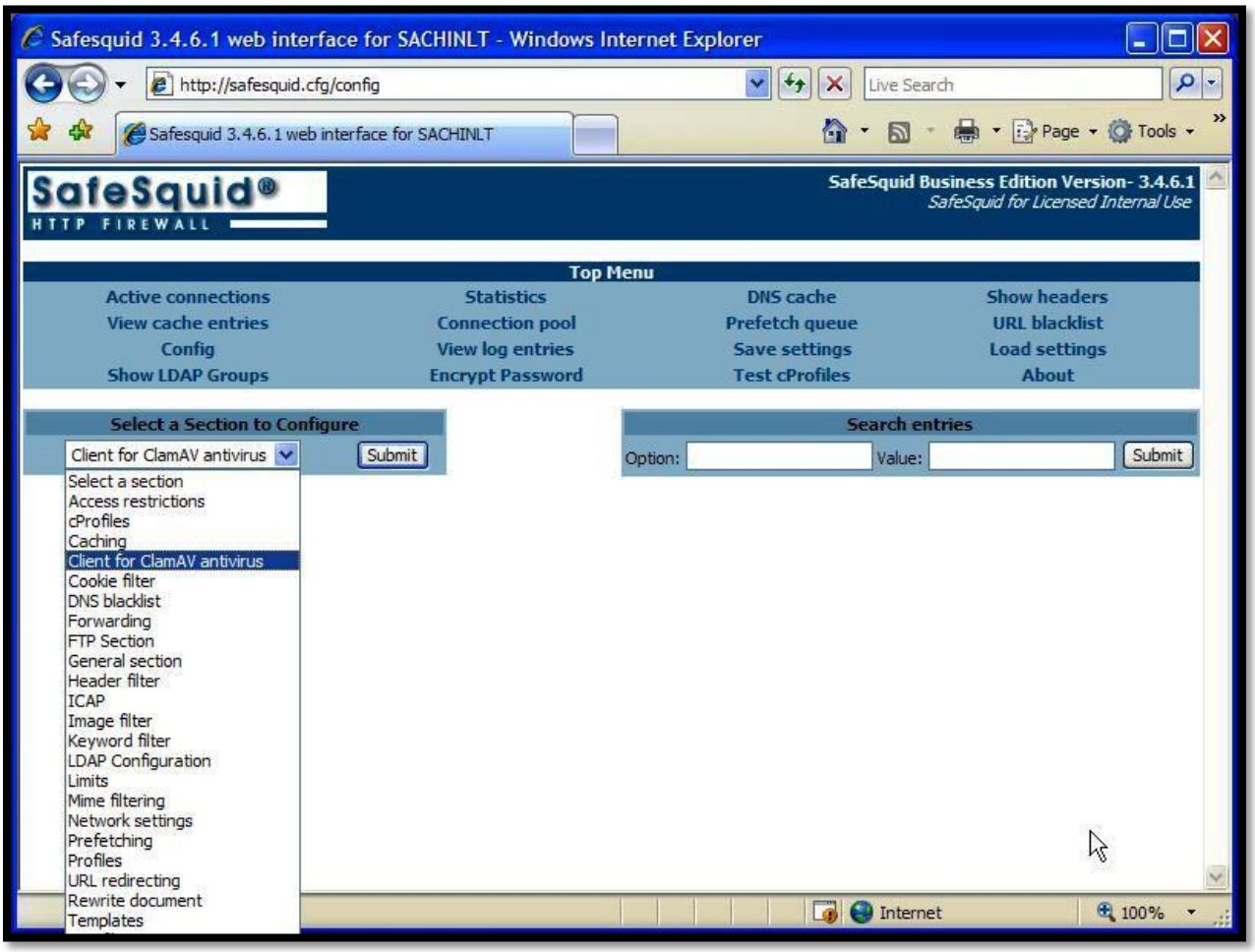
1.7 SquidGuar

- هو إضافة add-on على squid؛
- يقوم بعمل حجب للمواقع التي لا تريد أن يتم تصفحها عبر أجهزة الشبكة؛
- يحتوي على قاعدة بيانات تحتوي على المواقع "السوداء"؛
- إمكانية عمل صلاحيات للمستخدمين بطريقة أفضل وأسهل من squid؛
- الحد من الوصول إلى شبكة الإنترنت لبعض المستخدمين إلى قائمة خوادم الويب المعروفة والمقبولة و / أو عناوين URL فقط؛
- منع الوصول إلى بعض خوادم الويب المدرجة في القائمة السوداء و / أو عناوين URL لبعض المستخدمين؛
- منع الوصول إلى عناوين URL تحوي كلمات غير مرغوب بها لبعض المستخدمين؛
- إمكانية استخدام التعابير النظامية لحجب أجزاء محددة من صفحة الويب قبل عرضها؛
- فرض استخدام domain names / حظر استخدام عنوان IP في عناوين URL؛
- إعادة توجيه عناوين URL المحظورة إلى صفحة أخرى؛
- إعادة توجيه مستخدم غير مسجل لاستمارة التسجيل؛
- إعادة توجيه الشرائط الإعلانية إلى صورة فارغة؛
- إمكانية توليد ملفات log files؛
- لديها قواعد وصول اعتمادا على الوقت من اليوم، يوم من الأسبوع، والتاريخ ...؛

- لديها قواعد مختلفة لمختلف مجموعات المستخدمين (Profiles/Grouping)؛
- مجاني، سريع، سهل التنصيب، سهل الاستخدام ومفتوح المصدر؛
- من أنظمة التشغيل التي تعمل معها:

- **AIX: 4.1.3, 4.3.2.0/egcs-2.91.66**
- **Dec-Unix: OSF1-4.0/gcc-2.7.2.3, 3.2C/gcc-2.7.2.3**
- **FreeBSD 4.x-STABLE gcc 2.95.3**
- **Linux: RedHat-5.2/gcc-2.8.1 RedHat-7.x/gcc-2.8.1, Gentoo 1.12.6/gcc-3.3.6**
- **Solaris: 2.6/gcc-2.7.2.3 2.6/gcc-2.95.3, 2.8/gcc-2.95.3**
- **CentOS: 4.4**

1.8 SafeSquid



الشكل 16 صورة توضح واجهة safe-squid

- هو إضافة add-on على squid؛
- يقوم بعمل حجب للمواقع التي لا تريد أن يتم تصفحها عبر أجهزة الشبكة؛
- يحتوي على قاعدة بيانات تحتوي على المواقع "السوداء"؛
- توفر خدمة ال caching؛

- توفر خدمة مراقبة المستخدمين؛
- ضبط عرض الحزمة bandwidth؛
- مستخدم من قبل المدارس والشركات ومزودي خدمة الإنترنت؛
- منتج برمجي تجاري، مجاني فقط في حال لدينا ثلاث مستخدمين أو أقل ومفتوح المصدر؛
- يؤمن واجهة تخاطبية لتسهيل عملية الاستخدام من قبل المدير ليقوم بوضع إعداداته الخاصة؛
- إمكانية حجب مواقع كاملة أو حجب الموقع بسبب احتواء ال URL الخاص فيه على كلمات محددة مسبقا؛
- إمكانية حجب المواقع اعتمادا على كلمات محددة؛
- إمكانية حجب أنواع محددة من الملفات مثل ملفات الصوت والصورة؛
- إمكانية استخدام التعابير النظامية لحجب أجزاء محددة من صفحة الويب قبل عرضها؛
- إمكانية إخفاء الشرائط والصفحات الإعلانية لتخفيف عرض الحزمة وبالتالي تحسين الأداء؛
- إمكانية حجب الصور السيئة الغير المرغوب فيها من خلال تحليل الصور والتعرف على محتواها؛
- يولد ملفات logs لمراقبة المستخدمين ومن ثم ترجمة هذه الملفات وتوليد تقارير تحليلية؛
- يملك خاصية الملف الشخصي والذي يتيح للمدير أن يصنف المستخدمين ضمن مجموعات ووضع إعدادات الفلترة على مستوى المجموعات؛
- دقتها تراوح بين 85%-90% وتعتبر رادع جيد؛
- فضلا عن أنها تعتبر مضاد للفيروسات والبرمجيات الخبيثة؛
- أنظمة التشغيل التي تعمل معها Microsoft Windows, Linux.[7]

1.9 QuintoLabs

- مجاني ومفتوح المصدر؛
- يدعم أنظمة التشغيل التالية:

Linux distributions (RedHat, CentOS, Fedora, Debian, Ubuntu), Windows

- يعمل كإضافة على مخدم البروكسي Squid؛[5]
- يقوم QuintoLabs بالفلترة بواسطة عدة تقنيات هي:

1- يقوم بضمان فلترة المواقع المُزارة (من مواقع البالغين) من خلال تقنية RTL Labeling وهي عملية وضع علامات خاصة متل HTTP headers أو meta بالصفحات والصور التابعة لمواقع البالغين، تتم هذه العملية من قبل الموقع طوعاً، فعندما يقوم QuintoLabs بفحص الصفحات يبحث عن هذه العلامات فإذا وجدها يقوم بحجب الصفحة، تقنية RTL ليس لها أي إعدادت خاصة وإنما فقط يمكن تفعيلها أوإلغاء تفعيلها on/off؛

2- URL-Heuristics يتم استخدام هذا module لحجب URLs معينة التي غالباً ما تكون مرتبطة بمواقع البالغين، وهنا يوجد عدة مستويات من الصرامة في الاستدلال تتبعها الخوارزمية مثلا heuristics_level = high، إذا كان مستوى الصرامة عالي جداً فإننا قد نحصل على بعض الإجابات الخاطئة بمعنى أنه قد نحصل على نتيجة بأن URL ما هو تابع لموقع له علاقة بالبالغين رغم أنه لا يكون فعلا كذلك!؛

3- Full Text Content Inspection وهي عملية الفحص الكامل للنص الموجود في الصفحات وأغراض HTML, Json، حيث يتم تحديد الكلمات والعبارات المراد حجبها ضمن ملفات محددة، وعندما

يقوم التطبيق بفحص صفحة ما يقوم بالبحث عن هذه الكلمات وحساب وزن الصفحة اعتماداً عليها فإذا تجاوز هذ الوزن حداً معيناً يتم حجب الصفحة، تعد هذه العملية مكلفة، لذلك تم توفير إمكانية تحديد الحجم الأقصى للملف الذي سيطبق عليه عملية الفحص الكامل لتجنب فحص الملفات الكبيرة التي سيأخذ فحصها وقتاً كبيراً، الحجم الافتراضي 300k؛

4- إمكانية حجب المواقع فوراً من خلال اسمه أو عنوانه URL وذلك من خلال كتابتها في ملفات خاصة:

مثلا حجب موقع حسب العنوان:

url = http://.*sex\.com.*

حجب موقع حسب الاسم:

domain = www.porno.com

حجب جميع المسارات الجزئية subdomains في الموقع السابق:

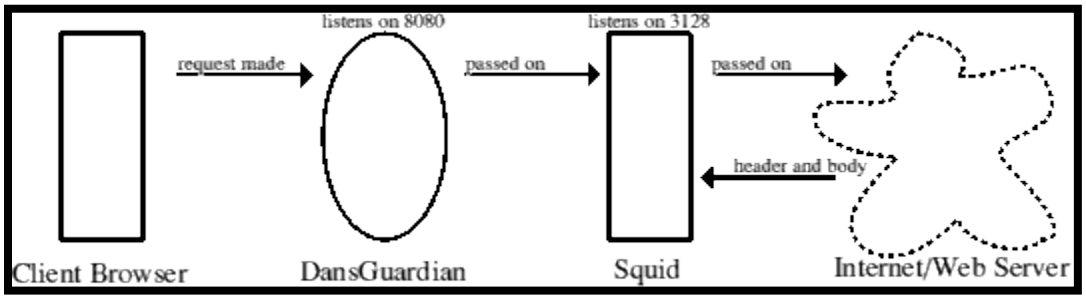
domain = .porno.com

كما يمكن حجب المواقع من خلال وضعها ضمن تصنفيات معينة ومن ثم حجب التصنيف، ويمكن إضافة التصنيف بالشكل التالي:

[6].block_category = sexuality

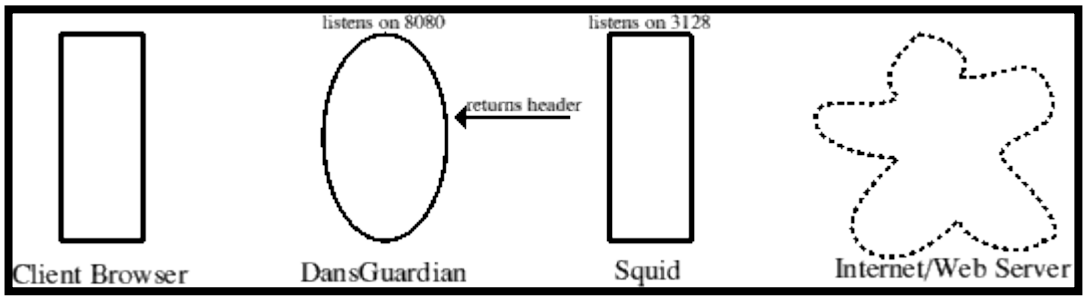
<u>مثال عن آلية عمل أحد الفلاتر وهو DansGuardian</u>

- يرسل المستخدم الطلب إلى DansGuardian، يقوم DansGuardian بفحص ترويسة الطلب القادم من المستخدم آخذا بعين الاعتبار اسم المستخدم، URL، IP المرسل؛
- يتم تطبيق الفلاتر المناسبة banned user, exception user, banned URL, exception URL, banned IP, exception IP
- في حال كانت النتيجة "قبول" يتم تمرير الطلب إلى البروكسي squid والذي يقوم بإحضار الملف المطلوب من الإنترنت؛



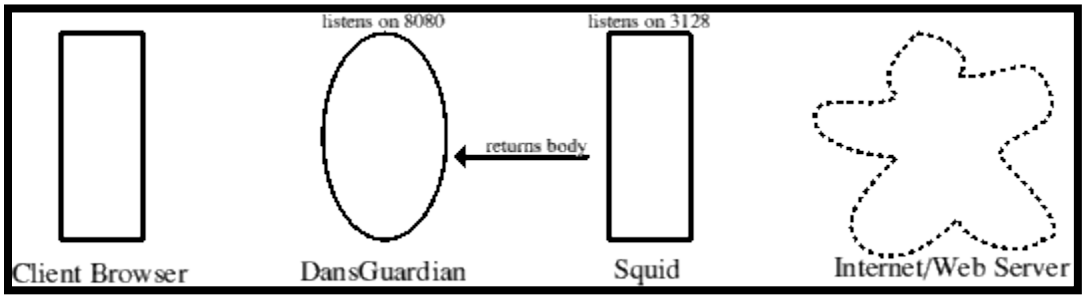
الشكل 17 آلية العمل DansGuardian + squid مرحلة 1

- يقوم squid يتمرير ترويسة الملف فقط إلى DansGuardian، والذي بدوره يقوم بفحص الترويسة بتطبيق مجموعة من الفلاتر المناسبة banned MIME-types وامتداد الملف؛
- في حال كانت النتيجة "قبول" يتم الانتقال إلى الخطوة التالية؛



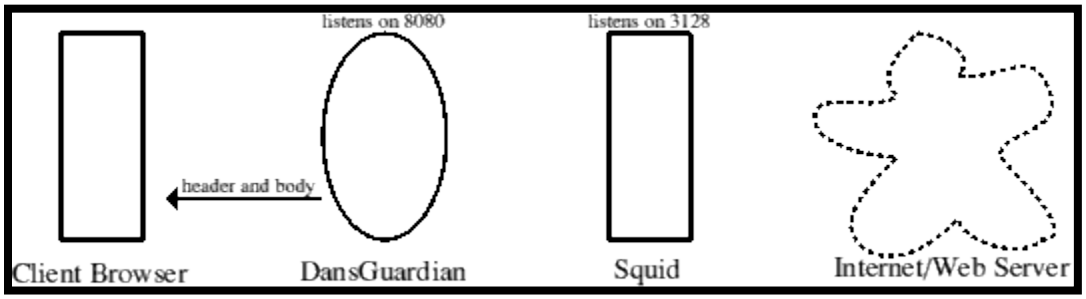
الشكل 18 آلية العمل DansGuardian + squid مرحلة 2

- يقوم squid بتمرير محتوى الملف أو الوثيقة إلى DansGuardian؛ والذي بدوره يقوم بتقسيمه إلى مجموعة عبارات ومن ثم يقوم بتطبيق الفلاتر المناسبة banned phrase, exception phrase؛



الشكل 19 آلية العمل DansGuardian + squid مرحلة 3

- في حال كانت النتيجة "قبول" يقوم DansGuardian بتمرير ترويسة ومحتوى الملف إلى متصفح المستخدم الطالب.



الشكل 20 آلية العمل DansGuardian + squid مرحلة 4

- تصويت من قبل موقع http://lifehacker.com

Which Content Filtering Tool is Best?

- **OpenDNS 57%**
-
- **K9 12%**
- **Hosts File 10%**
- **DansGuardian 8%**
- **Other 8%**
- **SquidGuard/Squid 5%**

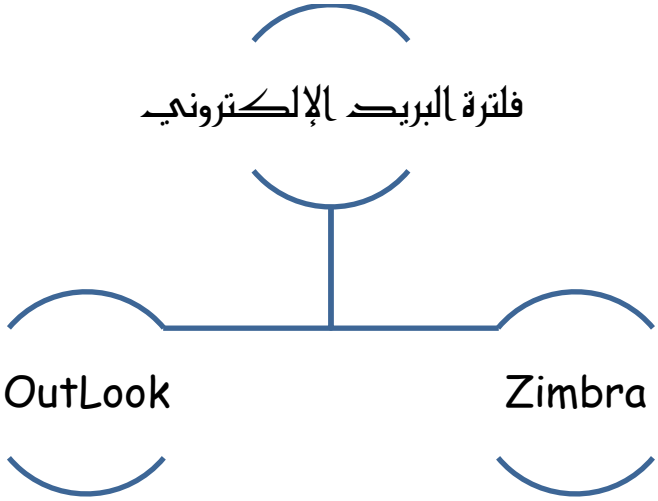
1.1.1 Hosts File

- ملف شبيه بدليل الهاتف يحوي المواقع المطلوب حجبها مع ال **ips** المقابلة لها؛
- عند محاولة الدخول إلى أحدى هذه المواقع عن طريق كتابة ال **URL** أو ال **IP** المقابل وكان ذلك العنوان موجود ضمن الملف على أنه موقع محجوب سيتم إظهار رسالة خطأ وإلا سيتم تحميل الصفحة الموافقة للعنوان المطلوب.

الفصل الثالث

دراسة مرجعية عن فلترة وتصنيف البريد الإلكتروني

نعرض في هذا الفصل دراسة مرجعية عن كيفية إدارة outlook و برنامج zimbra للبريد الإلكتروني، وآلية إضافة القواعد المطبقة على البريد الإلكتروني عند وروده بغية حجبه أو وضعه في مجلد خاص وغير ذلك من العمليات التي من الممكن تطبيقها على البريد الإلكتروني الوارد أو المرسل.



الشكل 21 النقاط الاساسية في فلترة البريد الإلكتروني

1 إدارة Outlook للبريد الإلكتروني من خلال القواعد

- القاعدة هي حدث يتم تطبيقه على البريد الإلكتروني الوارد أو المرسل اعتمادا على شروط محددة تم تعريفها مسبقا ضمن القاعدة.
- يمكن اختيار أكثر من شرط ضمن القاعدة وهذا ما يسمى Meta Rules.
- تتوزع القواعد ضمن مجموعتين مجموعة تدعى Organization ومجموعة تدعى Notification.
- هناك أنواع للقواعد:

1- Stay Organized فمثلا يمكن وضع قاعدة للرسائل الواردة من مرسل معين مثل Bobby Moore والتي تحوي الكلمة sales ضمن العنوان وعند ورود رسالة تحقق هذه القاعدة يتم وضعها في مجلد Bobby's Sales .

2- Stay Up to Date هذه القواعد تنبه المستقبل عند وصول رسالة خاصة فمثلا يمكن وضع قاعدة لإرسال تنبيه على الموبايل عند ورود بريد إلكتروني من أحد أفراد العائلة.

3- Start from a blank rule يتم بناء القواعد هنا من دون الاستعانة ب rule template..

- Outlook يتضمن templates للقواعد، يمكن استخدام هذه النماذج لبناء القواعد أو أن يتم تصميم القواعد بطريقة واضع القواعد.

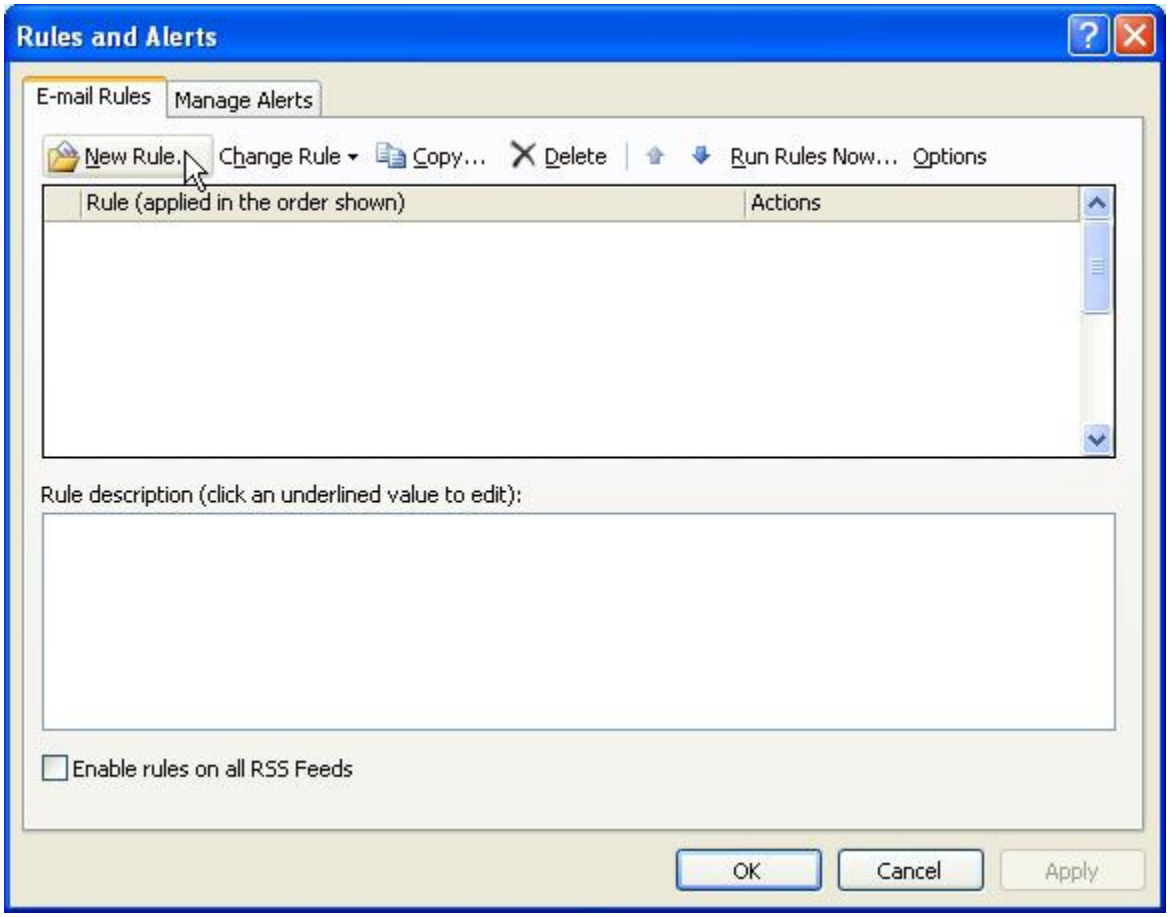
1.10 استخدام Outlook rule

1- يتم اختيار القائمة file؛

2- يتم اختيار القائمة Manage Rules & Alerts؛

الشكل 22 واجهة outlook مرحلة

3- من Rules and Alerts يتم اختيار القائمة E-mail Rules ومن ثم اختيار New Rule؛

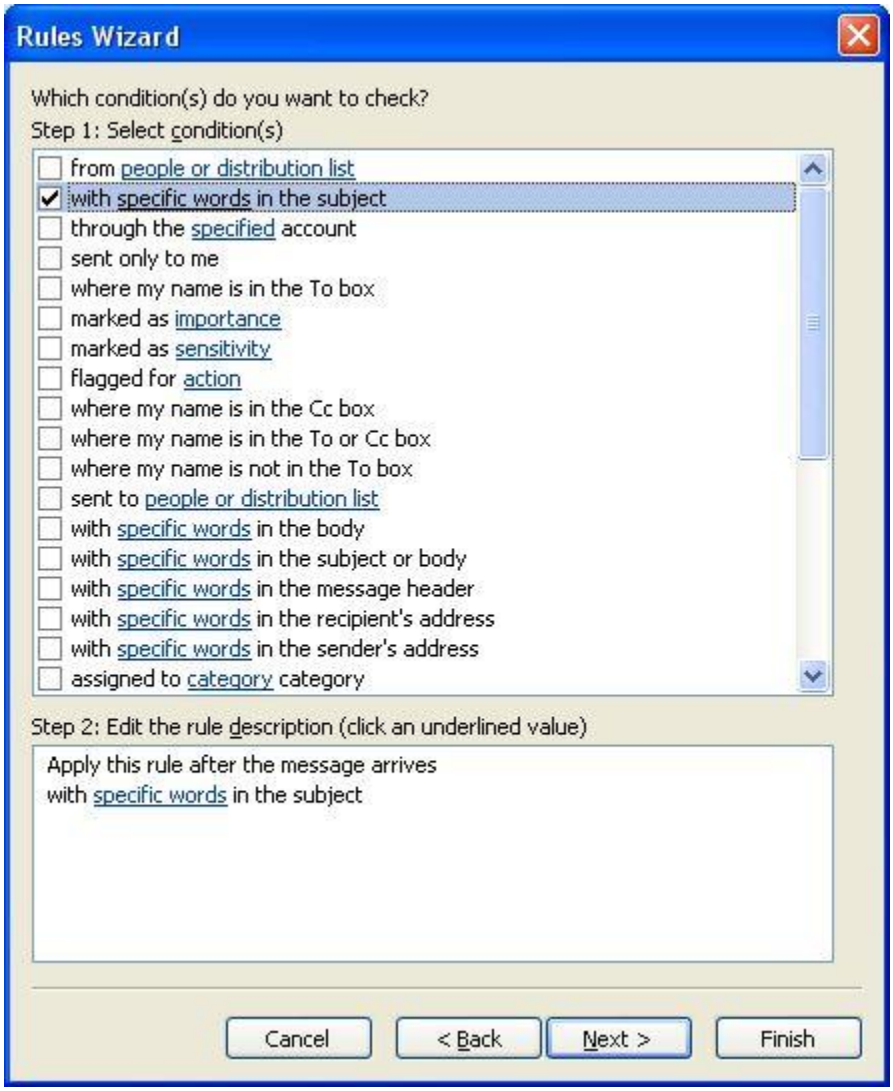


الشكل 23 واجهة outlook مرحلة 2

4- بعد ذلك تظهر نافذة Rules Wizard وتختار قالب محدد أو بناء القاعدة من دون الاعتماد على القوالب الموجودة مسبقا، وبفرض أنه تم اختيار قالب Start from a blank rule وتحديدا فحص البريد الإلكتروني عند وردوده، هذا ويمكن تعديل توصيف القاعدة.

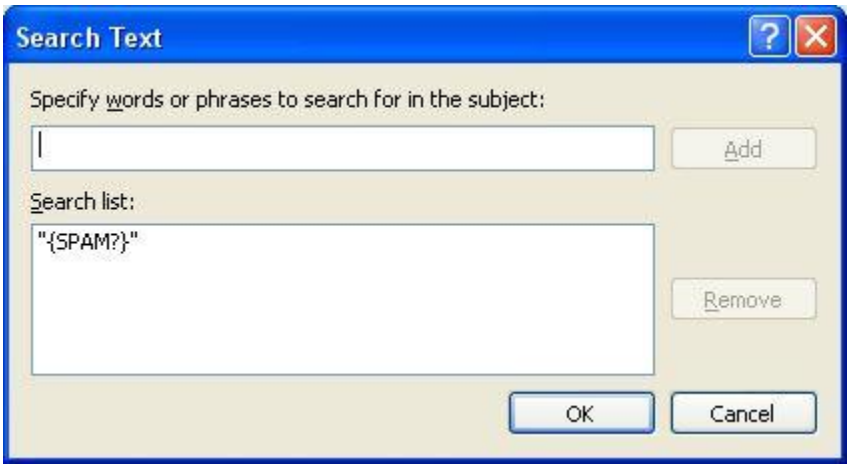
الشكل 24 واجهة outlook مرحلة 3

5- يتم اختيار مجموعة الشروط التي يتم تطبيقها على الرسالة على سبيل المثال فلترة البريد الإلكتروني بناء على كلمات محددة في موضوع الرسالةن كما يمكن تعديل التوصيف كما هو مبين في الشكل.

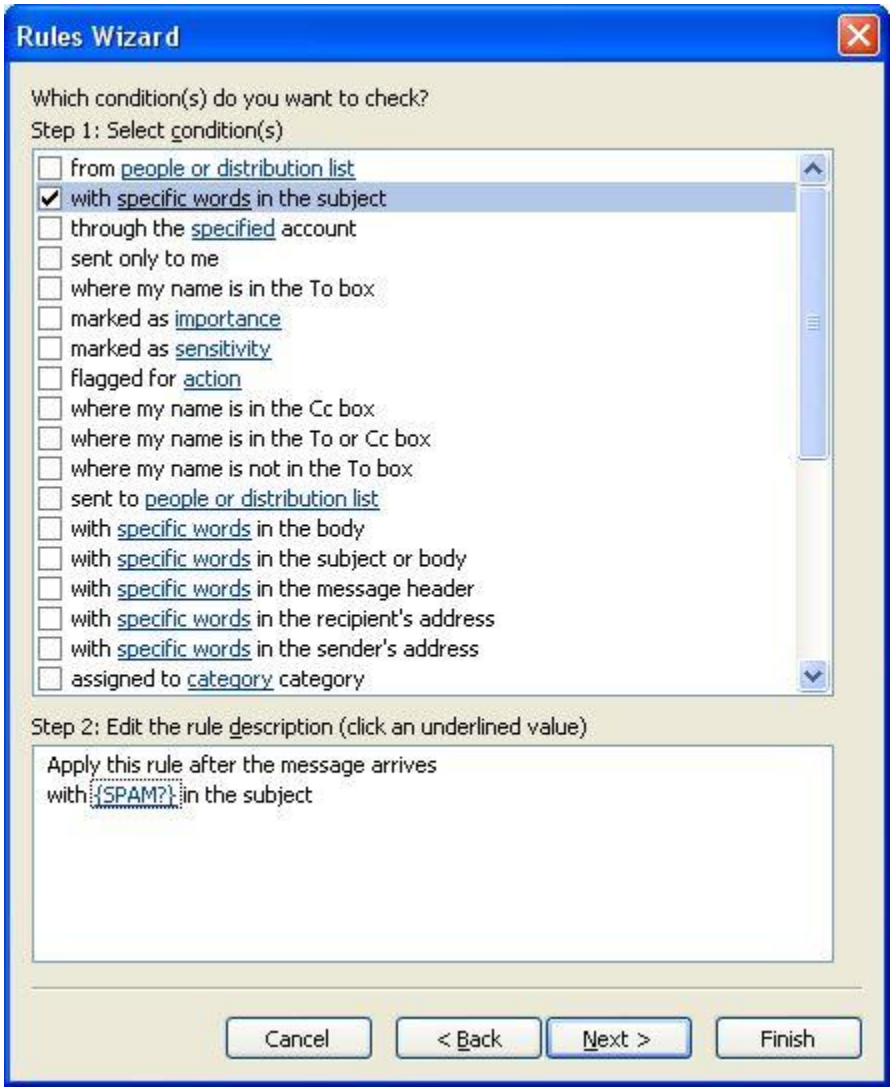


الشكل 25 واجهة outlook مرحلة 4

6- تحديد الكلمات الغير مرغوب بها والتي ستتم الفلترة على أساسها

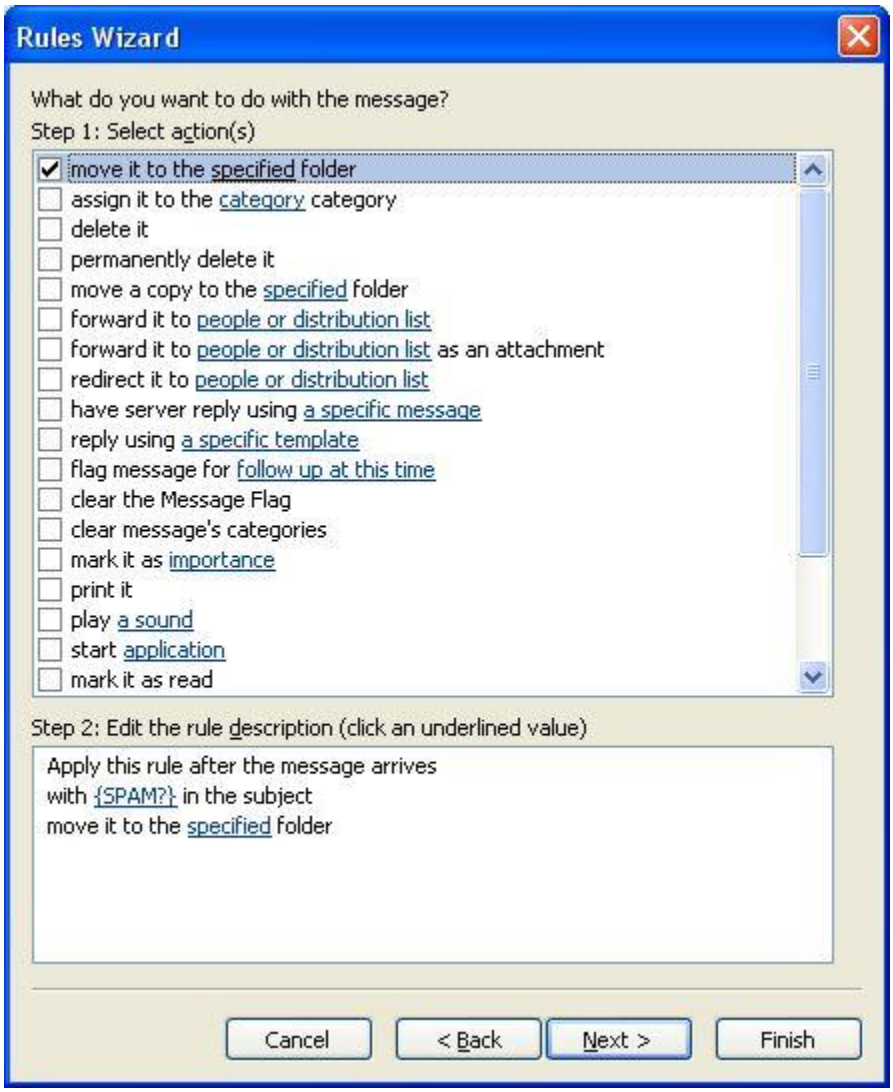


الشكل 26 واجهة outlook مرحلة 5



الشكل 27 واجهة outlook مرحلة 6

7- نقل البريد الإلكتروني الذي يحقق هذه القاعدة إلى مجلد خاص وليكن spam أو تفعيل خيار حذف هذه الرسالة في حال ثبت أنها spam



الشكل 28 واجهة outlook مرحلة 7

8- تفعيل خيار إيقاف فحص بقية القواعد عند تحقق هذه القاعدة

الشكل 29 واجهة outlook مرحلة 8

2 Zimbra Anti-spam system

- Zimbra هو نظام Anti-spam يستخدم لفلترة البريد الإلكتروني؛
- الطريقة الأسهل هو التعديل على ملف /opt/zimbra/conf/salocal.cf.in؛
- الطريقة الأبسط المتبعة في الفلترة هي القوائم البيضاء والقوائم السوداء؛
- القوائم السوداء تحجب البريد الإلكتروني الوارد من عنوان أو domain محدد؛
- القوائم البيضاء تسمح بعرض البريد الإلكتروني الوارد من عنوان أو domain محدد؛
- لإضافة القوائم السوداء أو البيضاء يتم الدخول إلى الملف salocal.cf.in والكتابة بالصيغة التالية على سبيل المثال:

```
blacklist_from sales@traveloforange.com

whitelist_from bill@yahoo.net

blacklist_from *@emn-mysavingsnow.net
```

- حيث يتم حجب كل المستخدمين التابعين لل domain التالي emn-mysavingsnow.net؛
- عند الانتهاء من التعديل على الملف يتم إعادة تشغيل zimbra من خلال التعليمة التالية:

zmmtactl restart && zmamavisdctl restart

- يتم قراءة عنوان ومحتوى البريد الإلكتروني وتطبيق القواعد عليهما؛
- القواعد هنا قد تحوي كلمات بعينها أو تعابير نظامية؛

- عندما يحقق البريد الإلكتروني القاعدة الموضوعة يتم زيادة الرصيد الكلي للبريد الإلكتروني بمقدار رصيد هذه القاعدة؛
- عندما يتجاوز الرصيد الكلي للبريد الإلكتروني عتبة محددة يتم اعتباره spam، وإذا كان أكبر بمقدار معين يتم حذفه مباشرة؛
- يتم إضافة هذه القواعد الحاوية على التعابير النظامية إلى الملف alocal.cf.in بالشكل التالي:

```
body LOCAL_RULE     /sale/

score LOCAL_RULE    0.5
```

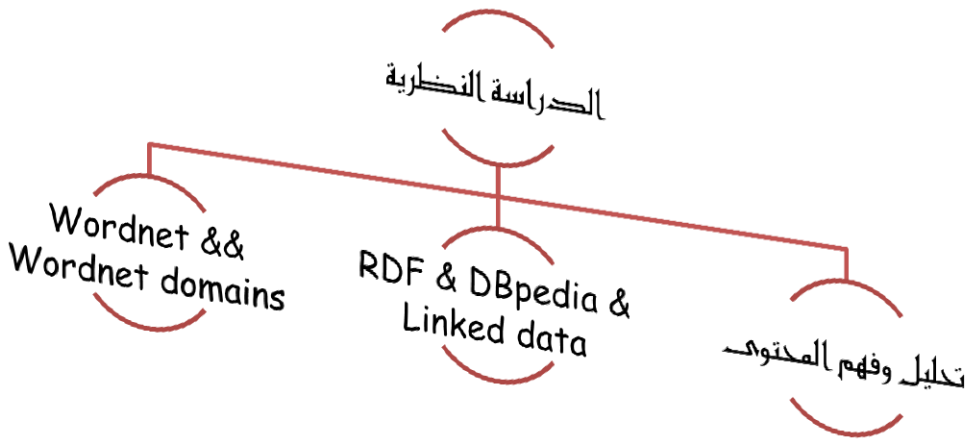
- القاعدة السابقة اسمها LOCAL_RULE تعني إذا احتوى البريد الإلكتروني على كلمة sale سيتم زيادة الرصيد الكلي للبريد الإلكتروني بمقدار 0.5؛
- Meta rule هي قاعدة تحوي على مجموعة قواعد يجب أن تتحقق معاً حتى يتم زيادة الرصيد الكلي للبريد الإلكتروني بمقدار معين مثال :

```
body LOCAL_FOUR_CAPS     /[A-Z][A-Z][A-Z][A-Z]/
body LOCAL_MONEY         /\d?\d?\d?.\d\d\b/
meta LOCAL_STOCK         (LOCAL_MONEY && LOCAL_FOUR_CAPS)
score LOCAL_STOCK        1
```

- بمعنى إذا حقق البريد الإلكتروني القاعدة LOCAL_FOUR_CAPS والقاعدة body LOCAL_MONEY سيتم زيادة الرصيد الكلي للبريد بمقدار 1.

الباب الثاني

الدراسة النظرية



الشكل 30 النقاط الأساسية في مرحلة الدراسة النظرية

الفصل الرابع

Wordnet && Wordnet Domains

نعرض في هذا الفصل لمحة عن معجم wordnet حيث تم الاعتماد عليه في فهم المحتوى، كما نعرض مجالات هذا المعجم والتي يتم الاعتماد عليها في تحديد أهم المجالات التني تتحدث عنها الصفحة ونسبة انتماء كل مجال إلى الصفحة.

1 معجم wordnet3

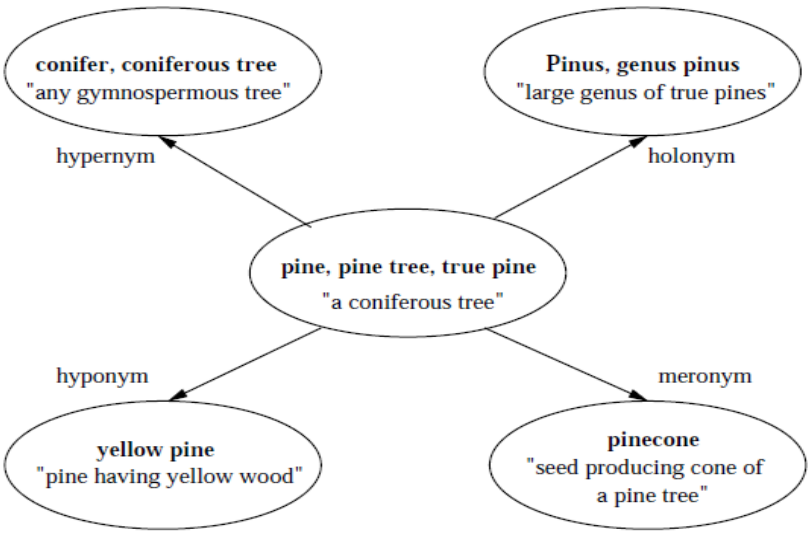
معجم الـ WordNet عبارة عن معجم يضم كلمات اللغة الانكليزية، بالإضافة إلى شرح هذه الكلمات المجمعة في مجموعات (Synsets)، بحيث أن كلمات كل مجموعة لها نفس المعنى (synonyms)، و هناك مجموعة من العلاقات الدلالية (semantic) التي تربط هذه المجموعات .

من المعلوم أن بعض الكلمات لها معنى وحيد (monosemous) و الكثير من الكلمات تحمل أكثر من معنى واحد (polysemous) وذلك حسب سياق الحديث، فعلى سبيل المثالـ كلمة " wristwatch " ليس لها إلا معنى وحيد، وبالتالي فهي تنتمي إلى مجموعة واحدة، في حين كلمة " bark" تحمل أكثر من معنى، فمن الممكن أن تأتي بمعنى لحاء الشجر، أو بمعنى صوت الكلب، وبالتالي فهي تنتمي إلى أكثر من مجموعة واحدة . [16]

في WordNet يتم وضع الكلمات المترادفة في مجموعات تدعى (synsets)، ولكل مجموعة تفسير (Gloss) وهو عبارة عن جملة تشرح المفهوم الذي تمثله هذه المجموعة، فعلى سبيل المثالـ المجموعة التالية : *basis*، *base*، { *foundation* } *fundament*، *groundwork*، *cornerstone* ، لها التفسير: "lowest support of a structure" .

وهذه المجموعات ترتبط مع بعضها البعض عن طريق علاقات دلالية وهي: hyponym ، hypernym ، meronym ، holonym .

مثالـ على العلاقات المختلفة التي تربط بين مجموعات الكلمات في الـ WordNet :



الشكل 31 مثال توضيحي عن العلاقات الدلالية في wordnet

أمثلة توضيحية عن العلاقات الدلالية المختلفة :

Hypernyms: Y is a hypernym of *X* if every *X* is a (kind of) *Y* (*canine is a hypernym of dog*. because every dog is a member of the larger category of canines).

Hyponyms: *Y* is a hyponym of *X* if every *Y* is a (kind of) X (*dog is a hyponym of canine*). --

Holonym: Y is a holonym of X if X is a part of Y (*building is a holonym of window*).

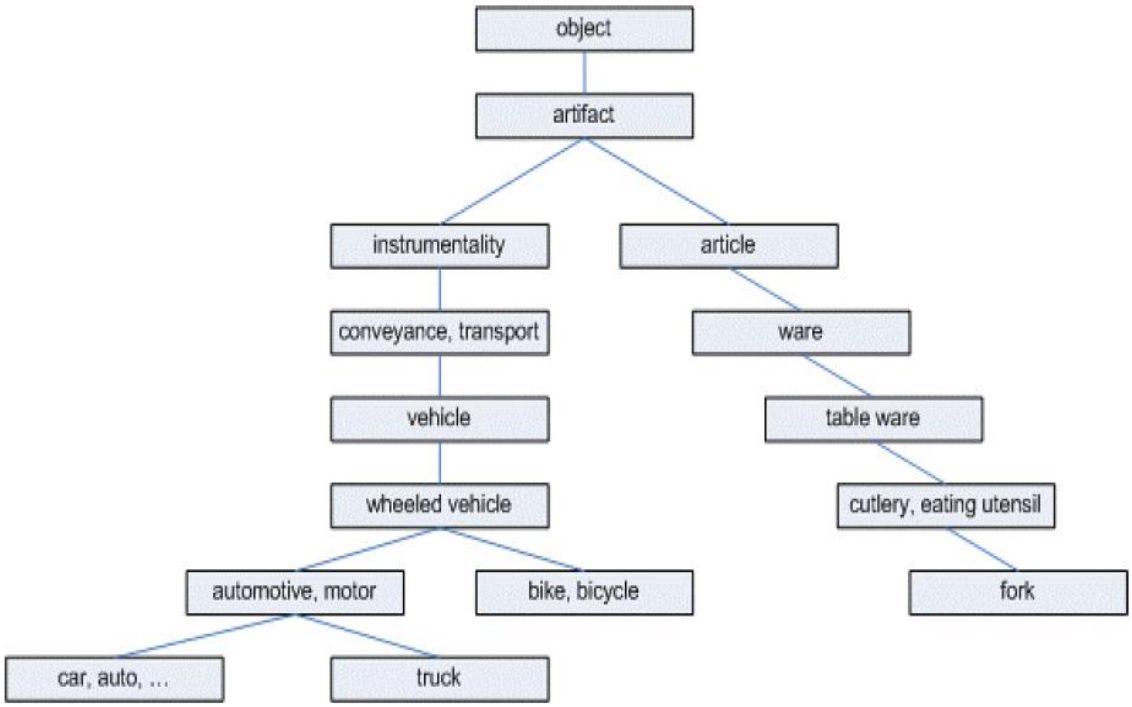
Meronym: Y is a meronym of X if Y is a part of X (*window is a meronym of building*).

Hypernym: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (*to perceive is an hypernym of to listen*).

Troponym: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (*to lisp is a troponym of to talk*).

ولكننا في دراستنا هذه سوف نهتم بالعلاقة hypernymy (is a kind of) و العلاقة المعاكسة hyponymy. كما أن للكلمات بنية هرمية باستخدام علاقتي hyponym و hypernym (is kind of) والتي نستطيع من خلالها إيجاد المفهوم (concept) لكلمة ما .

مثال: نلاحظ من الشكل التالي أن الـ "car" هي نوع من الـ "vehicle"،كما وتتيح هذه البنية معرفة مترادفات الكلمة car وهي auto . ،automobile ،machine ،motorcar



الشكل 32 مثال توضيحي عن الارتباطات في wordnet

من الجدير بالذكر أن الـ WordNet تقسم الكلمات إلى أربعة أنواع هي أسماء وأفعال وصفات وأحوال، وتصنف الكلمات تبعاً للأنواع السابقة وتوضع في مجموعات (synsets) حسب المعنى، مع العلم أن العلاقات السابقة الذكر بين المجموعات لا تربط إلا المجموعات التي تحوي الكلمات التي تنتمي إلى نفس النوع، فللأسماء المجموعات الخاصة بها وللأفعال مجموعات خاصة بها.

بعض الإحصائيات عن الـ WordNet :

POS	Words	Synsets	Senses
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Totals	155327	117597	207016

جدول2 إحصائيات عن معجم wordnet

سنوجه اهتمامنا في هذه الدراسة إلى الأسماء والصفات فقط لأنها تحمل معنى أوضح من الأفعال، فالفعل في أغلب الأحيان له معاني أكثر من الاسم، كما أن عدد الأسماء أكثر بكثير من عدد الأفعال.[8]

يمكن تحميل معجم الـ WordNet من خلال الرابط التالي : http://wordnet.princeton.edu/

2 جالات Extended Wordnet Domains

- Wordnet domains هو تصنيف لكل synset ضمن ال wordnet بالنسبة لكل ال domains المتاحة وكم نسبة انتماء هذه ال synset إلى كل domain؛
- يوجد حوالي 170 domain؛
- هناك 5 abstract domains ويتفرع عنها باقي ال domains؛
- من أجل كل domain هناك ملف يحوي كل الكلمات الموجوة في ال wordnet مع نسبة انتماء هذه الكلمة إلى هذا ال domain.[9]

TOP LEVEL
--> doctrines
--> free_time
--> applied_science
--> pure_science
--> social_science
--> factotum

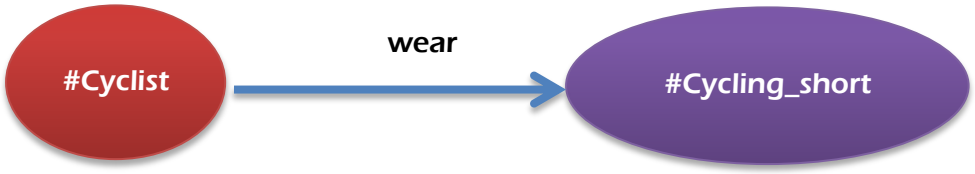
الفصل الخامس

RDF && DBpedia && Linked Data

نعرض في هذا الفصل أنطلوجية dbpedia المستخدمة في فهم الكلمة في حال لم تكن الكلمة موجودة ضمن معجم wordnet، وهنا يتحقق مفهوم المعطيات المترابطة حيث تم ربط الكلمة بمعناها ضمن أنطولوجية dbpedia، سنعطي لمحة سريعة عن مفهوم rdf حيث يكون محتوى dbpediaعلى هيئة ثلاثيات rdf.

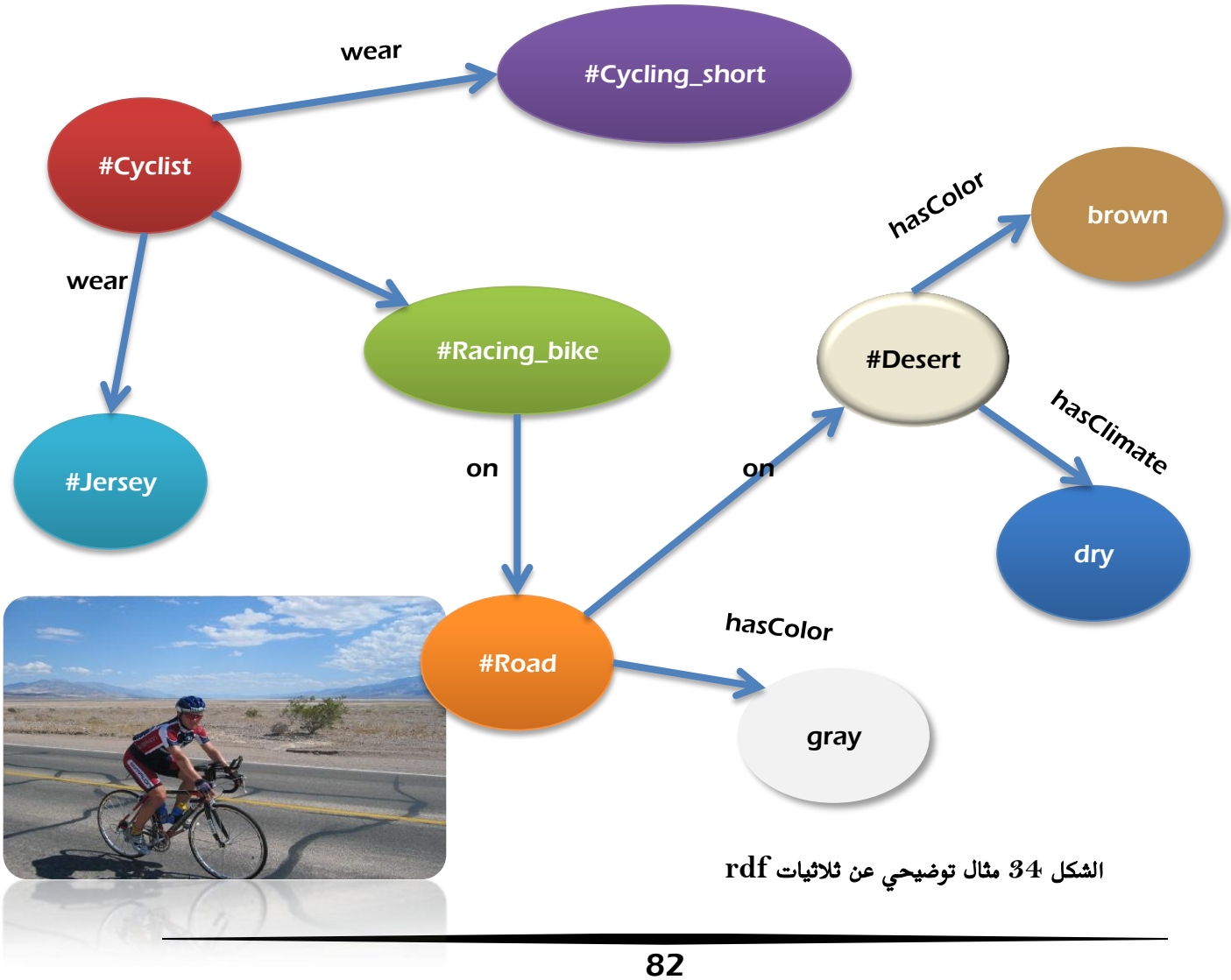
1 ثلاثيات RDF

- هو اختصار لـ (Resource Description Framework) وهي عبارة عن مجموعة من المواصفات التي وضعها اتحاد الشبكة العنكبوتية (W3C) لتعريف بنية تحتية مرنة خاصة بتنظيم وإدراة خصائص البيانات التي تسمى (metadata) في الشبكة العنكبوتية؛
- metadata هي عبارة عن معلومات وصفية عن طبيعة البيانات والوثائق مثل مصدرها، وحجمها و التنسيق الخاص بها وخصائص أخرى خاصة بالبيانات؛
- إذاً فـ RDF قد تم تصميمها لتوفير إطار يعتمد على لغة XML التي يمكنها أن تقوم بتوحيد عملية تبادل خصائص البيانات بين التطبيقات المختلفة أو ما يسمى بخصائص المحتويات (metacontent)؛
- ثلاثيات ال RDF تتألف من Subject, Predicate and Object؛



الشكل 33 مثال عن ثلاثية rdf

- ومن الاستخدامات المحتملة للـ RDF محركات البحث، وأنظمة تقييم المحتويات، ومجالات أخرى تهتم بخصائص البيانات المتبادلة؛
- مثال يوضح ثلاثيات ال RDF.



الشكل 34 مثال توضيحي عن ثلاثيات rdf

2 DBpedia

- قاعدة معطيات مجانية وضخمة تأخذ البيانات المهيكلة من Wikipedia وتقوم بتحويلها إلى ثلاثيات RDF؛
- مقالات Wikipedia تحوي نصوص بطريقة مهيكلة مضمنة ضمن المقالة، يتم استحصال هذه المعلومات المهيكلة ووضعها في بينات dataset بحيث نستطيع الاستعلام عنها؛
- إحصائيات عن dbpedia أيلول 2011 تقول أنها تصف حوالي 3.64 مليون موضوع تتوزع هذه الإحصائيات كما في الجدول التالي:

شخص	**416,000**
مكان	**526,000**
ألبوم موسيقي	**106,000**
فيلم	**60,000**
لعبة فيديو	**17,500**
منظمة	**169,000**
مرض	**5,400**
لغة	**97**
رابط إلى صور	**2,724,000**
رابط لصفحات ويب خارجية	**6,300,000**
رابط إلى بينات rdf dataset	**6,200,000**

جدول 3 إحصائيات عن أنطولوجية DBpedia

- Dbpedia إذاً يستخدم ثلاثيات rdf في تمثيل البيانات، إحصائيات أيلول 2011 بأنه يوجد حوالي 1 billion من ثلاثيات ال rdf موزعة بين 385 million ثلاثية مأخوذة باللغة الإنكليزية و 665 million مأخوذة بلغات أخرى؛
- لغة الاستعلام المستخدمة في الاستعلامات عن معلومات موجودة ضمن dbpedia هي SPARQL.

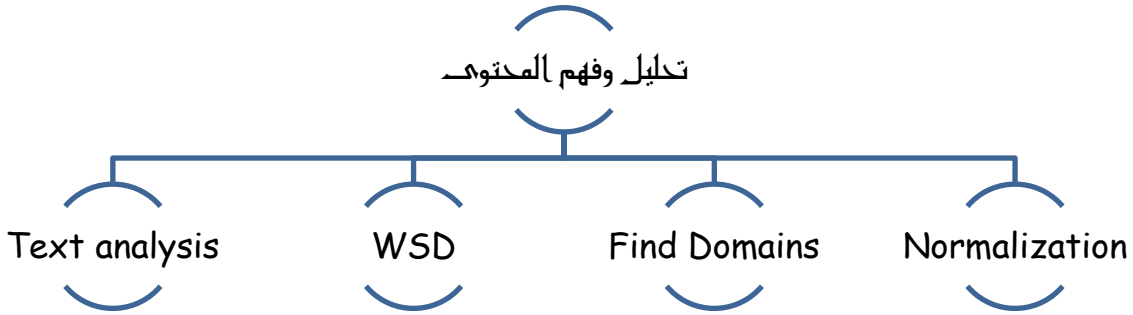
3 Linked Data

- يتم الربط بين مصادر مختلفة للبيانات مع بعضها البعض والاستفادة منها؛
- تم تطبيق هذه الفكرة في مشروعنا من خلال الربط مع dbpedia؛
- استخدمنا لأجل ذلك أداة تقدم لنا خدمة الربط مع dbpedia وهي spotlight؛
- في حال لم تكن الكلمة التي نهدف إلى فهم معناها ضمن wordnet يتم البحث عن الكلمة المطلوبة ضمن dbpedia؛
- يتم الاستفادة من العلاقات الموجودة في dbpedia ليتم فهم هذه العلاقات مثل علاقة ال subject، يتم فهم قيم هذه العلاقات من خلال تطبيق عمليات المعالجة الدلالية على هذه العلاقات وفهم الكلمة المطلوبة؛
- بهذه الطريقة نكون قد ربطنا المعطيات الموجودة لدينا مع شرح عنها ضمن أنطولوجية ال dbpedia .

الفصل السادس

تحليل وفهم المحتوى

نعرض في هذا الفصل الهدف الأساسي من المشروع والذي يقدمة خدمة تحليل وفهم المحتوى، عملنا على فهم المحتوى من خلال تطبيق خطوات متتالية من عمليات المعالجة الطبيعية ومن ثم الحصول على معاني الكلمات من خلال معجم ال wordnet، في حال لم تكن الكلمة موجودة يتم البحث عنها ضمن أنطولوجية dbpedia، بعد ذلك يتم معرفة المجال الذي تنتمي له كل كلمة وزيادة نقاط المجال بمقدار انتماء الكلمة إلى هذا المجال، وفي النهاية نقوم بعملية normalization للأوزان بحيث نحصل على نسبة انتماء كل مجال إلى الصفة التي نعمل على فهم وتحليل محتواها.



الشكل 35 النقاط الأساسية في تحليل وفهم المحتوى

1 استبدال الضمائر بالكلمات الأصلية التي تعود عليها هذه الضمائر Coreferencing

أحد المشاكل التي واجهتنا أثناء معالجة النصوص، هي ورود اسم شخص مثلا في بداية الجملة، ثم عند ذكر أفعال أخرى قام بها نفس الشخص فإنه يتم ذكر ضمير عائد على اسم الشخص بدلا من ذكر اسمه، و هذا يمنعنا من معرفة الفاعل الحقيقي لهذا الفعل مثلا:

John played football, and **he** also played golf.

نلاحظ أن الضمير he يعود على الفاعل John، لذا نقوم باستبدال الضمير بالفاعل قبل البدء بمعالجة النص وتحويله الى ثلاثيات فيصبح :

John played football, and **John** also played golf.

قمنا بالاستعانة بخدمة **Coreference** الموجودة ضمن مكتبة **Stanford CoreNLP**، لكننا قمنا بتطبيق بعض التحسينات عليها لمعالجة بعض المشاكل مثلا اذا أدخلنا المثال التالي:

Damascus is **the capital of Syria**.

ستكون النتيجة أن جملة the capital of Syria كلها تعود على Damascus، أي ستصبح الجملة بعد تطبيق العملية عليها:

Damascus is **Damascus**.

وهذه النتيجة غير مجدية أبدا، لذا قمنا بحصر هذه العملية فقط على الضمائر بكافة أنواعها (he, she, it, they...)، وإضافة 's في حال كان الضمير هو أحد ضمائر الملكية مثل(his, her ...).

كمثال على ذلك الجملة التالية التي تملك ضمير الملكية "his":

John likes all people. So people like **his** personality.

تصبح بعد المعالجة:

John likes all people. So people like **John's** personality.

2 إعادة الكلمات إلى أصلها Lemmatization

لمزيد من التحسين على النتائج قمنا أيضا بإعادة الكلمات إلى أصلها، حيث قمنا في البداية بتطبيق التجذير stemming على الكلمات لكن استنتجنا أن هذه الطريقة تحوي على العديد من المشاكل، فمثلا كلمة providers تتحول بتطبيق خوارزمية porter–وهي أحد خوارزميات التجذير– إلى كلمة provid

حيث نلاحظ أنها قامت بإزالة حرف ال e أيضا من نهاية الكلمة و بالتالي لم تعد الكلمة صحيحة لغويا ولا توجد في أي معجم.

لذلك قمنا باستخدام عملية أخرى تدعى lemmatization حيث تقوم هذه الطريقة بإرجاع الكلمة إلى أصلها لكن بالاستعانة بمعجم مثل WordNet وبالتالي نتائجها تكون صحيحة و موجودة في المعجم.

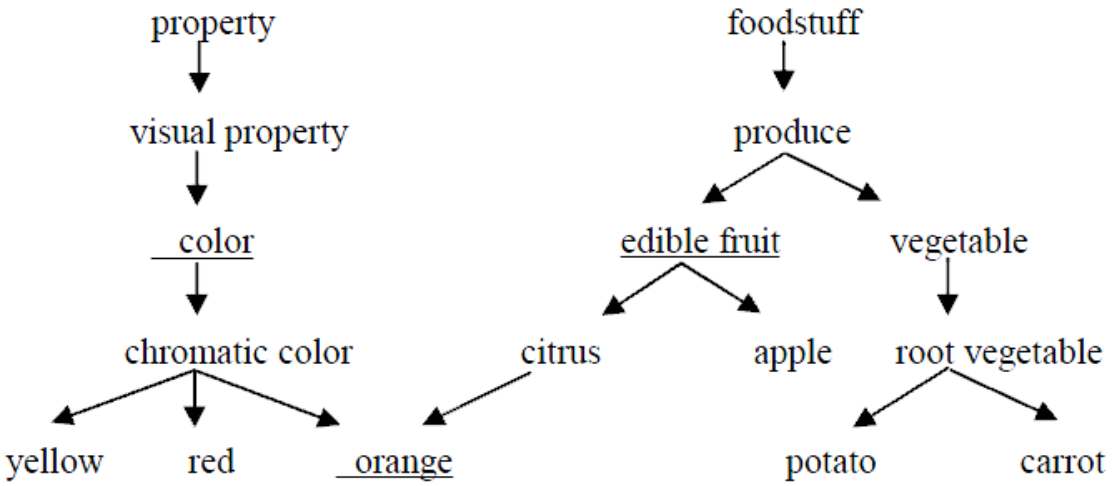
مثلا كلمة providers بعد تطبيق العملية عليها تعود الى كلمة provider.

3 إزالة غموض الكلمات Word Sense Disambiguation

يوجد في جميع اللغات الطبيعية كلمات يمكن أن يكون لها معنى مختلف في سياق مختلف (أكثر من معنى للكلمة الواحدة) . فك غموض الكلمة هي العملية التي تهدف لاكتشاف المعنى الصحيح لكلمة موجودة ضمن جملة بشكل آلي.

فعلى سبيل المثال، لليكن لدينا الشكل التالي:

نلاحظ أن للكلمة

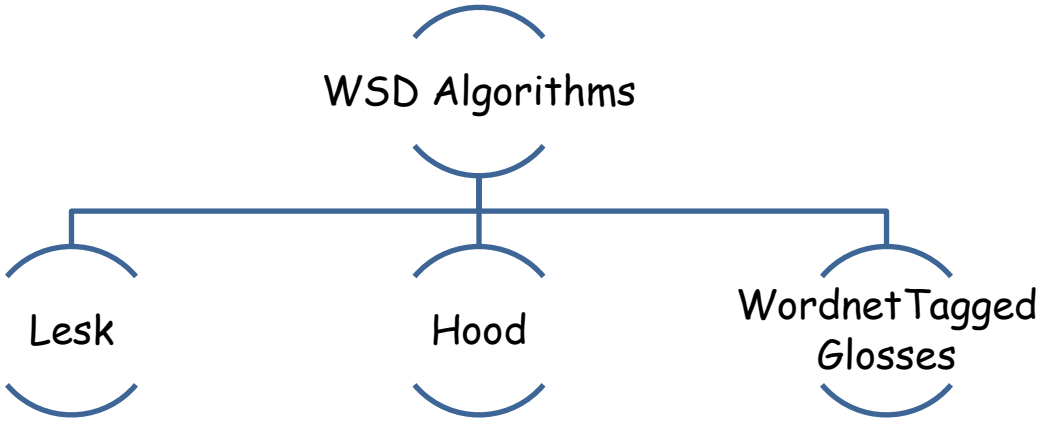


An example of two hypernym trees for the term orange.

الشكل 36 مثال عن غموض كلمة

"orange" شجرتي hypernym فيمكن أن تأتي بمعنى فاكهة أو لون، وبالتالي فلن نستطيع تحديد المفهوم المقابل لهذه الكلمة، وبالتالي يجب علينا أن نجد المعنى الصحيح للكلمة. هنالك العديد من الطرق لتحديد معنى كل كلمة منها ما هو بسيط، ومنها ما هو معقد، والنتائج تتفاوت في الدقة والسرعة.

فعلى سبيل المثال يمكن أن نعتبر أن المعنى الأول لكل كلمة هو المعنى الصحيح، وذلك باعتبار أن معاني الكلمات في الـ WordNet مرتبة تنازلياً وذلك حسب تردد استخدام الكلمة بهذه المعاني، فكلمة "dark" تستخدم بمعنى "مظلم" أكثر بكثير من استخدامها بمعنى "مغلق". كما و يمكن استخدام خوارزميات أكثر تعقيد مثل Lesk أو Hood أو غيرها.



الشكل 37 خوارزميات إزالة الغموض

3.1 خورازمية Lesk التقليدية

هذه الخوارزمية مبنية على الافتراضين التاليين :

1- عندما تستخدم كلمتين متجاورتين في جملة ما، فمن المؤكد إنهما يتكلمان عن نفس الموضوع.

2- إذا كان معنى واحد من كل كلمة يمكن أن يستخدم للتكلم عن نفس الموضوع، فلابد أن يكون تعريف المعجم لكل كلمة يستخدم كلمات مشتركة بينه وبين التعريف الآخر.

وبناءاً على هذين الافتراضين فإن خوارزمية تعمل على فك غموض الكلمات الموجودة ضمن جمل قصيرة (Short Phrases)، فمن أجل كل كلمة ضمن الجملة تقوم بإيجاد جميع تعريفات (Glosses) المعاني الخاصة بها (Sences) وذلك بالاستعانة بمعجم لكلمات اللغة الإنكليزية، ومن ثم تقوم بمقارنة كل تعريف من هذه التعاريف مع جميع تعاريف معاني بقية الكلمات في الموجودة في الجملة، إن عملية المقارنة بين كل تعريفين تتم بإيجاد عدد الكلمات المشتركة بينهما، ومنه فإن التعاريف المشتركة بأكبر عدد كلمات تكون هية التعاريف المناسبة لكلمات الجملة.

كمثال على ذلك نأخذ الجملة القصيرة التي تحوي الكلمتين الكلمتين ("pine - cone") فنجد أن هاتين الكلمتين فعلاً يتواجدان معاً وهما يتكلمان عن ("evergreen trees") و بالفعل نجد أنه يوجد في تعريف احدى المعاني لكل كلمة في المعجم الكلمتين ("evergreen", "tree"):

حيث نجد أن للكلمة الأولى pine معنيين هما:

Sense 1: kind of evergreen tree with needle–shaped leaves

Sense 2: waste away through sorrow or illness.

ونجد أن للكلمة الثانية core ثلاثة معاني هي:

Sense 1: solid body which narrows to a point

Sense 2: something of this shape whether solid or hollow

Sense 3: fruit of certain evergreen tree

بمقارنة كل معنى من معاني الكلمة ال pine مع جميع معاني الكلمة ال core نجد الكلمتين evergreen tree هما الأكثر تكرارا وبالتالي المعنى الأول هو الأنسب للكلمة pine والمعنى الثالث هو الأنسب للكلمة core.

Pine#1 Ç Cone#1 = 0

Pine#2 Ç Cone#1 = 0

Pine#1 Ç Cone#2 = 1

Pine#2 Ç Cone#2 = 0

Pine#1 Ç Cone#3 = 2

Pine#2 Ç Cone#3 = 0

نلاحظ من خورازمية Lesk التقليدية أنه على الرغم من بساطتها إلا أنها تعاني من مشكلة أن تعريف المعجم قصير جداً، ولا يوجد فيه كلمات كثيرة أو كافية من أجل أن تعمل هذه الخوارزمية بشكل جيد، بالإضافة إلى أنها لاتستفيد من المعاني التي قامت بإسنادها للكلمة خلال عملية إزالة الغموض حيث تقوم بإعادة العمليات من جديد من أجل كل كلمة.

3.2 التحسينات على خورازمية Lesk من خلال Global Disambiguation Using Glosses of Related Synsets

لقد أتت خوارزمية Global Disambiguation بعديد من التحسينات على خوارزمية Lesk التقليدية وذلك ما أدى إلى تحسين واضح في دقة إزالة الغموض وفي السرعة، حيث تعمل هذه الخوارزمية على الاستفادة من الكلمات المحيطة بالكلمة لإزالة غموض الكلمة الحالية، حيث يتم تعريف نافذة ذات طول فردي (الطول الأنسب 3 أي الكلمة الهدف وكلمتين محيطتين واحدة على يمينها وواحدة على يسارها)، ومن التحسينات التي تم إجراءها:

3.2.1 استخدام شبكة دلالية

إن خوارزمية Lesk التقليدية كانت تعتمد على المعاجم للحصول على تعاريف معاني كلمات الجملة مثل معجم Oxford Advanced Learner's Dictionary، وكما ذكرنا أن هذه المعاجم تحوي على تعاريف قصيرة غالباً ما تكون غير مجدية عند استخدامها في خوارزمية Lesk، ولذلك تم اعتماد على الشبكات الدلالية لغناها بالعلاقات الدلالية التي ستفيدنا في الحصول على المزيد من التعاريف لمعاني الكلمات والذي بدوره سيؤدي إلى الحصول على المزيد من التقاطعات بين الكلمات ضمن التعاريف.

ومن أشهر الشبكات الدلالية هي Wordnet التي تحوي على العديد من العلاقات التي تربط الكلمات ببعضها، وتختلف هذه العلاقات باختلاف موقع الكلمة في الجملة، ومن أشهر هذه العلاقات علاقتي Hypernym و Hyponym والتي هية

مسؤولة عن ترابط الكلمات بهيكلية is-a والتي ستفيدنا بالحصول على المزيد من التعاريف للكلمة الواحدة من خلال تعاريف أباءها

Noun	Verb	Adjective
Hypernym	Hypernym	Attribute
Hyponym	Troponym	Also see
Holonym	Also see	Similar to
Meronym		Pertainym of
Attribute		

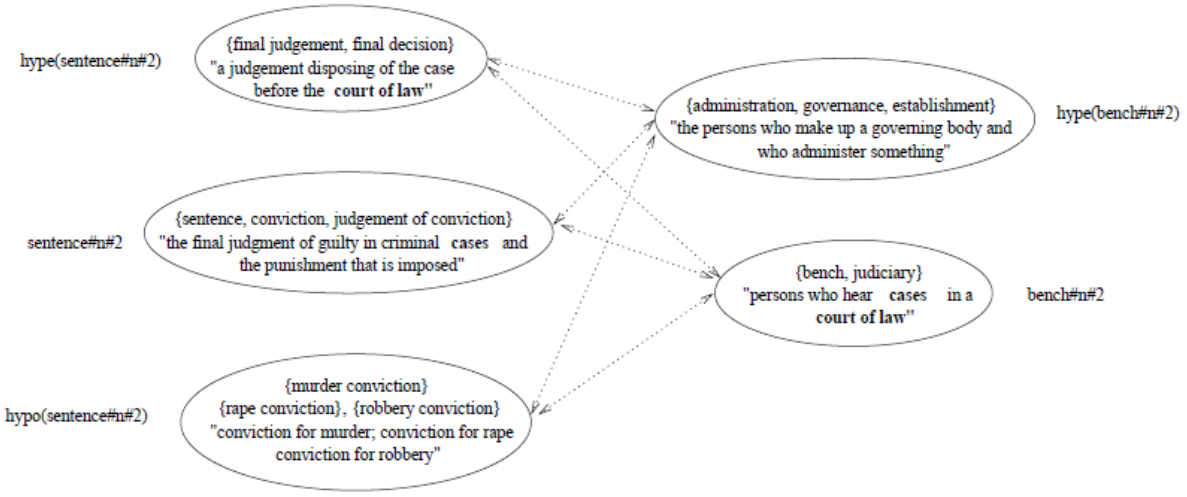
وأبنائها في الهيكلية.[17]

والجدول التالي يوضح أهم العلاقات التي من الممكن استخدامها مصنفة حسب موقع الكلمة من الكلام:

أساليب اختيار التعاريف الواجب مقارنتها

هناك عدة أساليب لاختيار التعاريف التي حصلنا عليها من الشبكة الدلالية التي يجب مقارنتها، فإذا اخذنا تعريفين لكلمتين وأوجدنا جميع التعاريف المرتبطة بهما بالعلاقات التي ذكرناها، فإنه من الممكن أن نقوم بمقارنة كل تعريف تابع لكلمة أو لإحدى علاقاتها مع التعاريف المرتبطة بالكلمة الأخرى ويدعى هذه الأسلوب Hetergeneous.

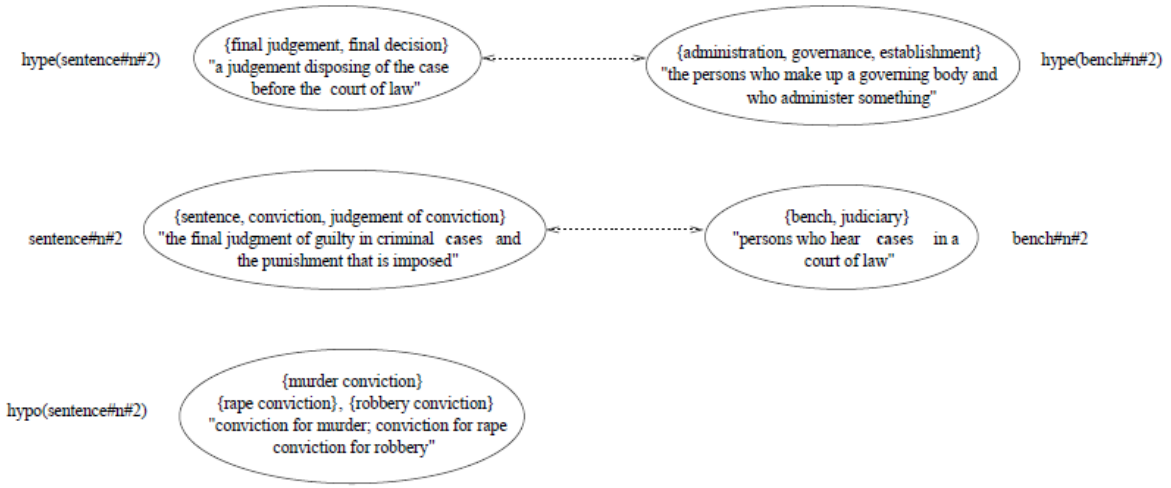
فإذا أخذنا الكلمتين Sentence و Bench وبأخذ العلاقاتين Hypernym المسماة hype و Hyponym المسماة hypo فإنه سيتم المقارنة بين الكلمتين وفق الشكل التالي:



الشكل 38 الدمج بين المجموعات

إن الكتابة sentence#n#2 ترمز إلى أن الكلمة sentence هي من النوع Noun وأن هذا هو المعنى الثاني لها.

كما بالإمكان اتباع أسلوب أخر في المقارنة حيث يتم فيه مقارنة التعاريف التي تنتمي إلى نفس الفئة فمثلا التعاريف التي تم إيجادها بالعلاقة hype في كلا الكلمتين يتم مقارنتها مع بعضها فقط، والتعاريف التي تم إيجهادها بالعلاقة hypo تتم مقارنتها سويةً فقط ويسمى هذا الأسلوب Homogeneous والشكل التالي يوضح اسلوب المقارنة هذا أكثر بين الكلمتين في المثال السابق:



الشكل 39 أسلوب المقارنة بين المجموعات

نلاحظ من الشكل السابق أنه لم نجد للكلمة bench أي رابط مع كلمات أخرى بعلاقة hypoولذلك فإنه لن يتم المقارنة عبر هذه العلاقة.

من الأسلوبين السابقين نلاحظ أننا في الأسلوب الأول Heterogeneous سنحصل على تقاطعات بين الكلمتين أكثر من الأسلوب الثاني Homogeneous ولكن في الأسلوب الثاني سنحصل على سرعة أكثر نظراً لقلة عدد المقارنات.[18]

3.2.2 ايجاد التقاطعات بين تعريفين

إن طريقة Lesk التقليدية كانت تعتمد على ايجاد عدد الكلمات المشتركة بين التعريفين، في هذه الطريقة أوجدنا أسلوب جديد لحساب التقاطعات بالاعتماد على أطول السلاسل المشتركة بين التعريفين ومن ثم توزينها وهذا الأسلوب يعتبر أكثر مجدياً من الطريقة التقليدية لإبرازه أهمية التقاطعات.

فلو كان لدينا التعريفين التاليين:

Sentence 1: you are the boy

Sentence 2: you are the boy I like

بالطريقة التقليدية نحصل على 4 كلمات متشابهة بين التعريفين وبالتالي سيكون الرصيد (score) هو 4 وهو نفس الرصيد الذي سنحصل عليه فيما اذا كانت الكلمات متباعدة عن بعدها ضمن كل تعريف، في حين أننا لو أخذنا المشترك هو أطول سلسلة كلمات مشتركة بين الجملتين فسنحصل على جملة واحدة مكونة من 4 كلمات، وفي حين كانت الكلمات متباعدة سنحصل على على 4 كلمات، وهنا نجد أن طريقة السلسلة المشتركة الأطول تأخذ بعين الاعتبار ورود الكلمات المشتركة مع بعضها بحيث تعطيها أهمية أكبر عن الكلمات فيما لو كانت متباعدة.

ولتحقيق هذه الأهمية لا بد من تربيع الرصيد الذي نحصل عليه من كل تشابه ، ففي المثال السابق يصبح الرصيد 4 * 4 = 16 هو رصيد الجملة، نلاحظ أنه أكبر من الرصيد الذي نحصل عليه فيما لو كان لدينا تشابهين أحداهما هو كلمة واحدة والآخر هو 3 كلمات حيث سيكون الرصيد هو (1 * 1) + (3 * 3) = 10 ، وأيضاً هذا الرصيد أكبر فيما لو كانت الكلمات الأربع متباعدة عن بعضها حيث سيكون (1 * 1) + (1 * 1) + (1 * 1) + (1 * 1) = 4. وبالتالي نجد كيف أن هذا الأسلوب فعلاً قد أعطى أهمية أكبر لورود الكلمات المتشابهة خلف بعضها عن الكلمات المتشابهة المتفرقة.

وقد تم استنتاج هذه الطريقة في الحساب من قانون الجبر التالي:

$$(a_0 + a_1 + \cdots + a_n)^2 > {a_0}^2 + {a_1}^2 + \cdots {a_n}^2$$

3.2.3 استراتيجية عمل خوارزمية Global Disambiguation

إحدى أبرز السيئات في خوارزمية Lesk التقليدية كما ذكرنا سابقاً هو أنها لاتستفيد من المعاني التي قامت بإسنادها للكلمة خلال عملية إزالة الغموض حيث تقوم بإعادة العمليات من جديد من أجل كل كلمة، وهذا ما تم معالجته أيضاً في هذه الخوارزمية حيث يتم تحديد نافذة من الكلمات المُراد معالجتها سويةً واستخراج المعاني الأنسب لجميع كلمات النافذة سويةً، أي أن كل كلمة في النافذة تساعد في إزالة غموض الكلمات الأخرى الموجودة ضمنها، وإن الحجم الأفضل للنافذة كما ذكرنا سابقاً هو 3 وذلك بسبب زيادة تعقيد العمليات الحسابية كلما زاد حجم النافذة أكثر.

بعد تحديد حجم النافذة نقوم بإيجاد جميع المعاني لكل كلمة من كلمات النافذة، ومن ثم إيجاد جميع التراكيب الممكنة (Combination) من معاني الكلمات مع بعضها، ومن ثم نقوم بإجراء مجموعة من عمليات المقارنة على كل تركيبة من المعاني وحساب رصيدها. في النهاية وبعد إيجاد رصيد كل التراكيب الممكنة نقوم باختيار اكبر رصيد موجود لتكون التركيبة الموافقة له هي التي تحمل المعاني الأنسب للكلمات الموجودة ضمن النافذة.

حساب رصيد التركيبة

1. نقوم بإيجاد جميع الأزواج الممكنة من الكلمات الموجودة ضمن كل تركيبة
2. من أجل كل زوج من الكلمات
   a. نقوم بإيجاد جميع العلاقات المرتبطة بكل كلمة في الزوج
   b. حسب أسلوب المقارنة المحدد (Heterogeneous, Homogeneous) نقوم بتحديد كيفية المقارنة للتعاريف التي تم ايجادها لكلمات الزوج والكلمات التي أوجدناها بالعلاقات
   c. نقوم بإيجاد التقاطعات بين كل تعريفين اعتمادا على الأسلوب الذي ذكرناه بالتحسين (أطول سلسلة مشتركة) لنحصل على رصيد نقوم بجمعه إلى رصيد التركيبة التي ينتمي إليها هذا الزوج.

ملاحظة: قد تكون المجموعة (synset) التي تحوي الكلمة الواحدة مرتبطة بعلاقة Hypernym (أو أي علاقة اخرى) مع عدة مجموعات أخرى (أي يكون لدينا عدة تعاريف)، عندئذٍ نقوم بجمع كل هذه التعاريف للمجموعات المرتبطة بنفس العلاقة بتعريف واحد يكون هو التعريف الممثل لارتباط الكلمة بهذه العلاقة مع بقية الكلمات.

مثال على تطبيق الخوارزمية:

لتكن لدينا نافذة ذات طول 3 مؤلفة من الكلمات التالية: sentence, offender, bench

بإيجاد معاني كل كلمة من الكلمات الثلاث نحصل على:

gloss(sentence#n#1) = a string of words satisfying the grammatical rules of a language

gloss(sentence#n#2) = the final judgment of guilty in criminal cases and the punishment that is imposed

gloss(bench#n#1) = a long seat for more than one person

gloss(bench#n#2) = persons who hear cases in a court of law

gloss(offender#n#1) = a person who transgresses law

بإيجاد جميع التراكيب الممكنة من هذه المعاني نحصل على المجموعة التالية:

Combination 1: sentence#n#1 – bench#n#1 – offender#n#1

Combination 2: sentence#n#1 – bench#n#2 – offender#n#1

Combination 3: sentence#n#2 – bench#n#1 – offender#n#1

Combination 4: sentence#n#2 – bench#n#2 – offender#n#1

الآن نقوم بأخذ كل تركيبة وإيجاد جميع الأزواج الممكنة وإجراء عملية المقارنة على كل زوج وحساب رصيد كل تركيبة، فمثلا إذا أخذنا التركيبة الأخيرة رقم 4 فإننا سنحصل على الأزواج التالية الموضحة بالجدول الآتي:

First Gloss	Second Gloss	Overlap String	Normalized Score
hype(sentence#n#2)	bench#n#2	court of law, case	10
sentence#n#2	bench#n#2	cases	1
hype(sentence#n#2)	offender#n#1	law	1
sentence#n#2	hypo(offender#n#1)	criminal	1
hype(bench#n#2)	hype(offender#n#1)	person	1
hype(bench#n#2	offender#n#1	person	1
hype(bench#n#2)	hypo(offender#n#1)	person	1
bench#n#2	hype(offender#n#1)	person	1
bench#n#2	offender#n#1	person, law	2
bench#n#2	hypo(offender#n#1)	person	1
Total score for sentence#n#2 – bench#n#2 – offender#n#1			20

الشكل 40 حساب رصيد كل تركيبة

نلاحظ من الجدول أن الأسلوب المتبّع في مقارنة الأزواج هو Heterogeneous.

وبالتالي نجد من الجدول أن رصيد التركيبة رقم 4 هو 20، وبالمثل نقوم بإيجاد رصيد كل التركيبات الأخرى واختيار الرصيد الأكبر لتكون التركيبة الموافقة هية الممثلة لكلمات النافذة.

تعقيد الخوارزمية

يحسب تعقيد الخوارزمية بالعلاقة التالية:

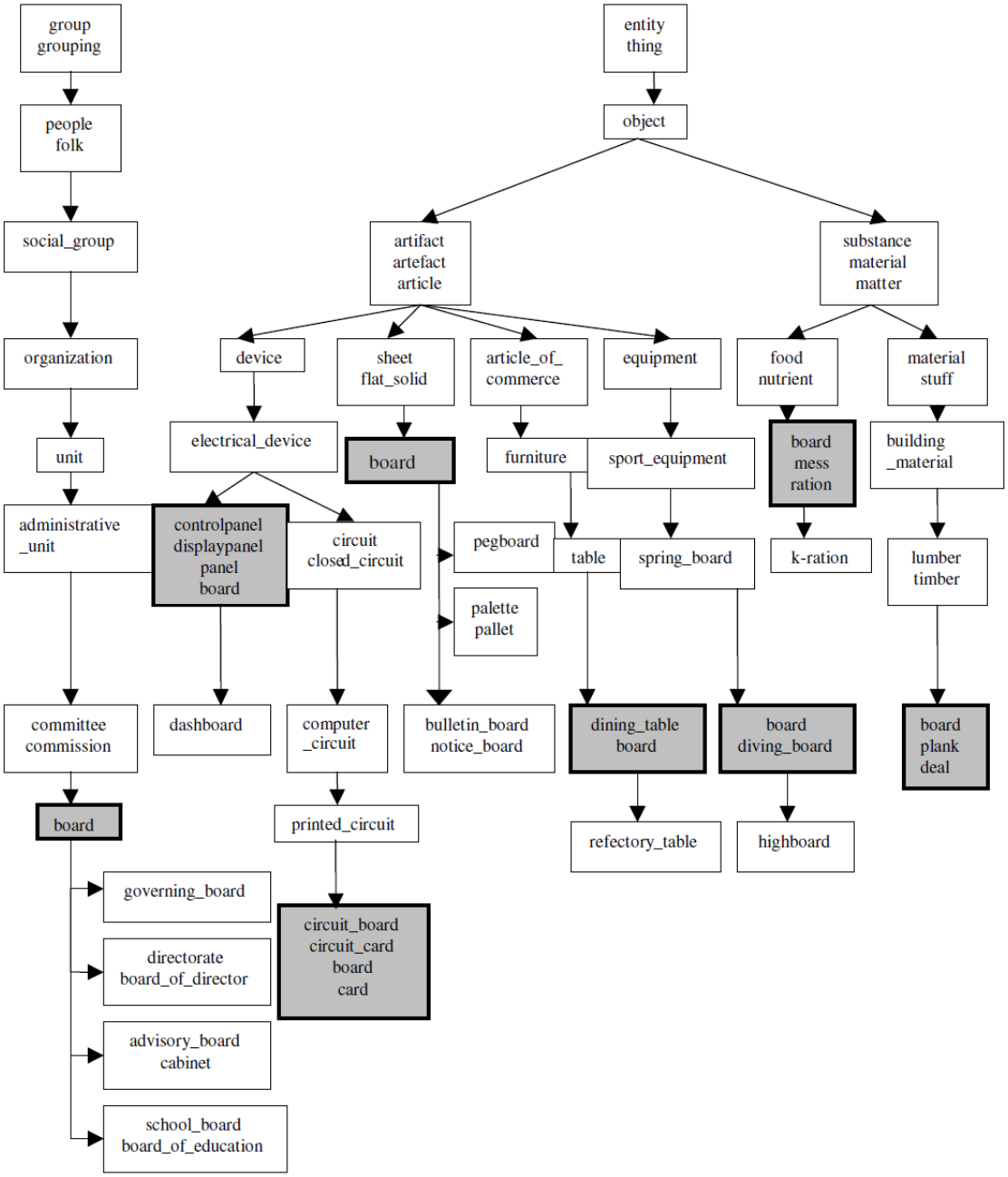
$$S^2 * \frac{N*(N-1)}{2}$$

حيث N: هو عدد الكلمات ضمن النافذة

S: هو وسطي عدد معاني كل كلمة

3.3 خوارزمية Hood

هي طريقة لتصنيف المستندات بشكل آلي مبنية على فك غموض الكلمات automatic text classification based on word sense disambiguation حيث يتم فك غموض مجموعة من الكلمات معا بنفس الوقت، وتبديل كل كلمة بالمعنى الصحيح لها، و هنا نجد اختلاف هذا الخوارزمية عن خوارزمية الـ primitive lesk وخوارزمية adaptive lesk و خوارزمية Weighted Overlapping، حيث أن هذه الخوارزميات تقوم بفك غموض الكلمات بشكل يدوي أو فردي manual (تأخذ الكلمات المراد فك غموضها كلمة كلمة حيث يتم فك غموض أول كلمة ثم التالية) أما خوارزمية Hood فيتم فك الغموض لمجموعة الكلام بنفس الوقت ، ويتم اعتبار أقرب سلف ancestor للمعنى الصحيح صف من صفوف تصنيف المستند ، مع العلم أن أغلب الأبحاث في word sense disambiguation تركز على استرجاع و الاستعلام عن المعلومات وليس على تصنيف المستندات (النصوص) .



الشكل 41 مثال عن إيجاد جذر hood

The IS-A hierarchy for eight different senses of the noun "board" with the root hood of each synset of the word "board"

3.3.1 فك غموض الكلمة وتصنيف المستندات word sense disambiguation and text classification

خوارزمية Hood مبنية على فكرة أن مجموعة من الكلمات المتواجدة معا سوف تحدد المعنى المناسب لكل كلمة منها ، على الرغم من أن كل كلمة لها أكثر من معنى .

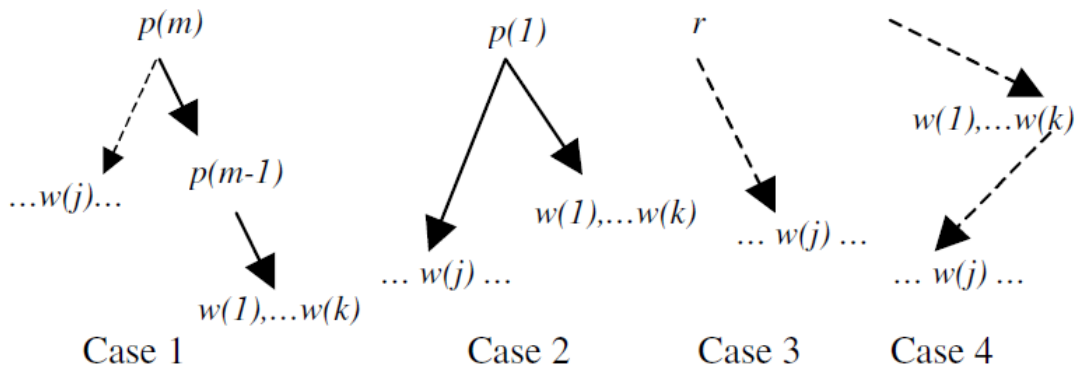
مثالـ : الكلمات التالية base،bat ،hit and glove كل كلمة منها لها أكثر من معنى و لكن عندما تكون معا فأن الموضوع هو لعبة البيسبول .

وبالتالي لاستغلالـ هذه الفكرة بشكل آلي، يجب تعريف مجموعة من الصفوف (classes) تمثل مختلف المعاني للكلمات ، ثم نأخذ عداد يقوم بعد الكلمات التي لها هذا المعنى و بالتالي المعنى الصحيح للكلمة يتحدد بالصف الذي له أكبر قيمة عداد و أقرب سلف لهذا المعنى يكون تصنيف للمستند.

3.3.2 إنشاء Hood (Hood Construction)

يعرف الـ Hood لمجموعة مترادفات synset بأنه أكبر بيان جزئي متصل يحوي synset ويحوي أبناء من أسلاف للـ synset ، و لا يحوي أية مجموعة مترادفات أخرى تملك ابن يملك كلمة موجودة في synset ، ويمثل الـ Hood بالجذر Hood Root .

حالات تحديد الجذر للـ Hood :



الشكل 42 حالات تحديد الجذر في خوارزمية hood

الحالة الأولى : أن يكون أحد أسلاف الـ synset يملك ابن يحوي كلمة من كلمات الـ synset و بالتالي الجذر هو السلف السابق لهذا السلف (case 1) .

الحالة الثانية : أن تكون الـ synset هي نفسها الجذر لأن السلف المباشر (الأب) لها يملك ابن يحوي كلمة من كلمات الـ synset (case 2) .

الحالة الثالثة : أن يكون جذر الهيكلية للـ WordNet هو الجذر اذا لم يوجد سلف يملك ابن يحوي كلمة من كلمات الـ synset مع عدم الإخلال بشرط تعريف الـ Hood (case 3) .

الحالة الرابعة : اذا كانت الـ synset نفسها تملك ابن يحوي كلمة من كلمات الـ synset بالتالي لا يوجد Hood لهذه الـ synset (case 4) .[19]

3.3.3 فك الغموض

- من أجل كل كلمة غامضة:
  - من أجل كل sense لهذه الكلمة يتم أخذ جذر hood المقابل:
    - نمرُّعلى جميع الـ synsets لجميع الكلمات المراد فك غموضها، و نأخذ علاقات الـ hypernyms لهذه الـ synsets ، وعند المرور Hood root يتم زيادة العداد لهذا الـ Hood Root قيمة واحد وبالنهاية يكون المعنى الصحيح للكلمة هو المعنى المتواجد ضمن synset وجذر hood المقابل له هو الجذر ذو العداد الأعلى.

3.4 إزالة الغموض بالاستعانة ب WordNet Semantically Tagged glosses

3.4.1 لمحة عن مشروع WordNet Semantically Tagged glosses

تعتمد هذه الطريقة على مشروع WordNet Semantically Tagged glosses والذي تم اطلاقه من قبل الجامعة المنتجة لأنطولوجيا WordNet، الهدف من هذا المشروع هو ازالة غموض جميع كلمات التعاريف glosses الخاصة باي معنى SynSet، اي انه يقوم بربط جميع كلمات التعريف بالمعنى الصحيح المقابل لها بالاستفادة من الكمات المحيطة.

فلنأخذ مثلا كلمة orangeونأخذ تعريف أحد المعاني الخاصة بهذه الكلمة:

Round yellow to orange fruit of any of several citrus trees

يهتم المشروع بإزالة غموض جميع الكلمات المظللة بشكل مسبق، و يقوم بربطها بالمعنى الصحيح لها ضمن WordNet.

نقوم بتسمية المعاني الناتجة عن ازالة غموض تعريف معنى معين بالمعاني المرتبطة بذلك المعنى، اي ان المعاني المرتبطة بمعنى orange هنا هي المعاني المناسبة لكل من round, fruit, citrus

وبالتالي يقدم لنا المشروع corpus ضخم جدا من النصوص التي تمت عمليه إزالة غموض كلماتها بشكل مسبق، ويمكننا الاستفادة منها بكثير من المجالات منها ازالة غموض نصوص أخرى.[14]

3.5 الخوارزمية المتبعة لإزالة الغموض :

تقوم فكرة الخوارزمية على انه عادة ما تأتي المعاني المرتبطة بمعنى معين ضمن الكلمات المحيطة بهذا المعنى، و هذا يساعد على معرفة المعنى الحقيقي للكلمة من بين جميع المعاني المحتملة

لفرض مثلا انه لدينا الجملة التالية و لنحاول معرفة المعنى الصحيح لكلمة **pen** ضمن هذه الجملة:

The **pen** is taking care of the small swans.

- نبحث عن المعاني المحتملة لكلمة pen ضمن WordNet فنجد انها يمكن ان تأخذ المعاني التالية:

1. a writing implement with a point from which ink flows
2. an enclosure for confining livestock
3. a portable enclosure in which babies may be left to play
4. a correctional institution for those convicted of major crimes
5. female swan

- نوجد الآن المعاني المرتبطة بكل معنى من هذه المعاني بالاستعانة WordNet Semantically Tagged glosses:

1. a writing implement with a point from which ink flows
    a. point
    b. ink
    c. writing implement
2. an enclosure for confining livestock

a. enclosure
b. livestock, stock, farm animal

3. a portable enclosure in which babies may be left to play
   a. portable
   b. enclosure
   c. baby, babe, infant
4. a correctional institution for those convicted of major crimes
   a. major
   b. crime, offense, criminal offense, criminal offence, offence, law-breaking
   c. correctional institution
5. female swan
   a. female
   b. swan

- نقارن المعاني المرتبطة بكل معنى من المعاني الخمسة لكلمة pen مع جميع المعاني المحتملة للكلمات المحيطة بكلمة pen في الجمالة المدخلة و نجمع 1 لل score الخاص بأحد المعاني الخمسة في حال حصل تطابق

- نلاحظ ان المعنى الخامس يأخذ ال score = 1 لتطابق المعنى swan من المعاني المرتبطة به مع احد معاني كلمة swan الموجودة في الجملة المدخلة، و المعاني الاربعة الأخرى تأخذ ال score = 0 لذلك يكون المعنى الحقيقي للكلمة هو المعنى الخامس

- في حال تساوى اكثر من معنى بال score وكان هذا اعلى scoreنأخذ المعنى الاكثر استخداما في اللغة (يمكننا الحصول عليه من WordNet) من بين هذين المعنيين.[15]

3.5.1 ايجابيات و سلبيات هذه الخوارزمية

تتميز هذه الخوارزمية بسرعتها العالية كونها تعتمد على علاقات محسوبة بشكل مسبق وتقوم فقط بعمليات مقارنة بسيطة، الا انها لا تقدم دقة عالية في النتائج، مما يجعلها مناسبة للتطبيقات التي تحتاج لسرعة عالية و لا تطلب دقة كبيرة في النتائج.

4 مقارنة بين الخوارزميات الأربعة المستخدمة في إزالة الغموض

اسم الخوارزمية	السرعة	الفعالية
خوارزمية المعنى الأكثر شهرة	★★★★	★
خوارزمية Wordnet tagged glosses	★★★	★★★★
خوارزمية Hood	★★	★★
خوارزمية Lesk	★	★★★

جدول4 مقارنات بين الخوارزميات الاربعة في إزالة الغموض

4.1 طريقة تحليل وفهم المحتوى اعتمادا على معالجة اللغات الطبيعية

4.1.1 قراءة صفحة الإنترنت المطلوبة وتحويلها إلى ملف نصي

استخلاص النص من صفحة الوب المطلوبة سواء كان text, doc, pdf, docx, html…..؛

4.1.2 تجزئة الكلمات

يتم تجزئة الكلمات التي تحوي على أحرف كبيرة في منتصفها أو التي تحوي المحرف "-" فذلك يعني أنها كلمة جاءت كتركيب أو أنها اسم علم؛

4.1.3 إزالة المحارف الغريبة

تتم عملية إزالة المحارف الغريبة من النص مثل المحارف التي لا تنتمي إلى الأبجدية الإنكليزية أو الأرقام لأن لا معنى لها ضمن معالجتنا للنصوص؛

4.1.4 تقسيم النص إلى جمل

يتم تقسيم النص إلى جمل حسب علامات الترقيم حيث يتم تمييز جملة عن أخرى من خلال المحرف "." وتمييز الكلمات ضمن الجمل ومعرفة موقع الكلمة ضمن الجملة هل هي فعل أم اسم أم صفة أم ظرف؛

4.1.5 المعالجة الدلالية وعملية الفلترة

- من أجل كل جملة من الجمل التي تم الحصول عليها من المرحلة السابقة
  - من أجل كل كلمة token ضمن الجملة؛
    - يتم استبدال الضمائر بالكلمات الأصلية التي تعود عليها هذه الضمائر Coreferencing وذلك ضمن خيار يتم تفعيله أو إلغاء تفعيله من قبل مستخدم النظام كون هذه العملية تستغرق وقتا؛
    - التحقق من موقع الكمة ضمن الكلام هل هو اسم أو صفة؛
    - التحقق من أن هذه الكلمة هي كلمة تحمل معنى أي أنها لا تنتمي إلى مجموعة ال stopwords؛

- التحقق من أن طول الكلمة يتجاوز المحرفين كون أن أي كلمة تتألف من محرفين فقط هي كلمة لا تحمل معنى ويجب الاستغناء عنها؛
- الحصول على أصل الكلمة من خلال عملية ال Lemmatization وذلك بهدف الحصول على كلمة موجودة ضمن معجم wordnet والاستغناء بذلك عن عملية ال stemming؛
- محاولة إزالة غموض هذه الكلمة والحصول على المعنى الحقيقي لها حسب سياق الجملة الواردة ضمنها؛
- إيجاد المجالات التي تنتمي لها الكلمة domain من خلال الاستعانة ب wordnet domains ومن ثم الحصول على وزن هذه الكلمة ضمن كل domain والانتقال إلى شجرة المجالات وزيادة مجموع كل domain بوزن الكلمة ضمن هذا ال domain وزيادة مجموع المجالات الآباء التي يتبع لها ال domain الحالي؛
- في حال لم تكن الكلمة موجودة ضمن معجم ال wordnet نحاول البحث عنه ضمن أنطولوجية dbpedia والاستفادة من علاقة ضمن هذه الأنطولوجية والتي تشرح معنى هذه الكلمة حيث يتم إعادة تطبيق نفس الخطوات من أجل شرح معنى هذه الكلمة ضمن dbpedia.

- نقوم بعملية normalization للأوزان التي حصلنا عليها بعد تطبيق المراحل السابقة ضمن شجرة wordnet domains tree حيث يتم إعطاء ال domain الحاصل على أعلى وزن القيمة 1 وإعادة توزين ال domains المتبقية اعتمادا على نسبة مئوية من ال domain الأعلى وزن.
- بعد ذلك يتم تمرير شجرة domains tree إلى مجموعة القواعد التي تم وضعها من قبل خبراء وبناء على رغبة مستخدمي هذا المنتج ليتم معالجة تلك القواعد اعتمادا على الشجرة الممررة ومعرفة هل سيتم حجب صفحة الوب المطلوبة أم لا.

4.2 طريقة فهم المحتوى اعتمادا على ثلاثيات RDF

1- استخلاص النص من صفحة الوب المطلوبة سواء كان text, doc, pdf, docx, html…..؛

2- تجزئة الكلمات التي تحوي على أحرف كبيرة في منتصفها أو التي تحوي المحرف "-" فذلك يعني أنها كلمة جاءت كتركيب أو أنها اسم علم؛

3- إزالة المحارف الغريبة من النص مثل المحارف التي لا تنتمي إلى الأبجدية الإنكليزية أو الأرقام لأن لا معنى لها ضمن معالجتنا للنصوص؛

4- تقسيم النص إلى جمل حسب علامات الترقيم حيث يتم تمييز جملة عن أخرى من خلال المحرف "." وتمييز الكلمات ضمن الجمل ومعرفة موقع الكلمة ضمن الجملة هل هي فعل أم اسم أم صفة أم ظرف؛

5- يتم تحويل النص إلى ثلاثيات subject, predicate and object؛

6- يتم ربط subject, predicate مع أنطولوجيات معروفة مثل dbpedia من خلال URIs هي صلة الوصل ما بين هذه الثلاثيات والأنطولوجيا؛

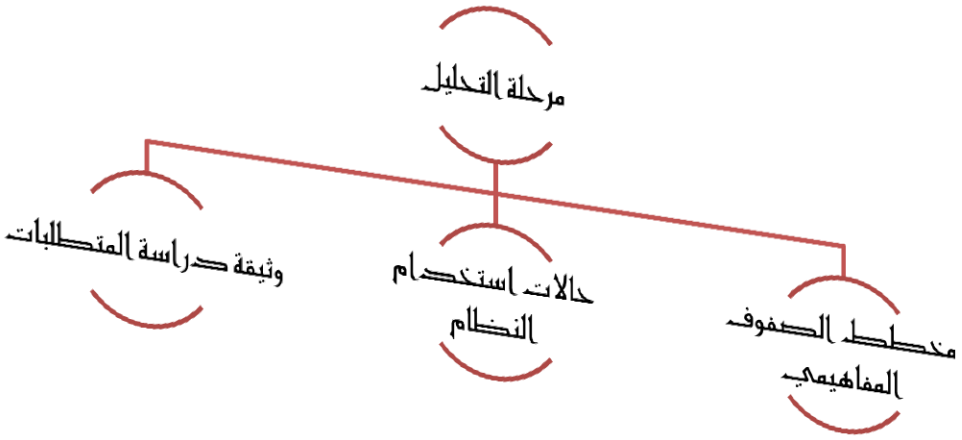
7- يتم فهم محتوى هذه الثلاثيات من خلال العلاقات الموجودة بالأنطولوجية ومحاولة معرفة ال wordnet domain الخاص بها من خلال هذه العلاقات ومن ثم الحصول على وزن هذه الكلمة ضمن كل domain والانتقال إلى شجرة المجالات وزيادة مجموع كل domain بوزن الكلمة ضمن هذا ال domain وزيادة مجموع المجالات الآباء التي يتبع لها ال domain الحالي؛

8- نقوم بعملية normalization للأوزان التي حصلنا عليها بعد تطبيق المراحل السابقة ضمن شجرة wordnet domains tree حيث يتم إعطاء ال domain الحاصل على أعلى وزن القيمة 1 وإعادة توزين ال domains المتبقية اعتمادا على نسبة مئوية من ال domain الأعلى وزن؛

9- بعد ذلك يتم تمرير شجرة domains tree إلى مجموعة القواعد التي تم وضعها من قبل خبراء وبناء على رغبة مستخدمي هذا المنتج ليتم معالجة تلك القواعد اعتمادا على الشجرة الممررة ومعرفة هل سيتم حجب صفحة الوب المطلوبة أم لا.

الباب الثالث

مرحلة التحليل



الشكل43 النقاط الأساسية ضمن مرحلة التحليل

الفصل السابع

وثيقة دراسة متطلبات النظام

وهي عبارة عن وثيقة تكتب من قبل الشركة المطورة للمنتج، توضح من خلالها مميزات المنتج الذي تقوم ببنائه، وتتعتبر وثيقة تفاهم بين الشركة المطورة للمنتج والشركات التي ترغب بالحصول عليه.

من خلال هذه الوثيقة يتم تحديد متطلبات المشروع ويتم الاستفادة من خبرات المهندسين من أجل تزويد المشروع بأفضل الحلول من أجل تحقيق هذه المتطلبات.

1 تعريف المشكلة المدروسة

تزايد كبير ومستمر في صفحات الإنترنت والمحتوى الضخم الموجود على الشبكة العنكبوتية، جعل من غير الممكن القيام بعمليات تحليل وفهم مسبق للمحتوى، عمليات المعالجة التي تطبق على المحتوى من فلترة وتصنيف وغير ذلك من غير العملي والمنطقي أن يتم تطبيقها بشكل مسبق.

يمكن القيام بهذه العمليات بشكل يدوي لكن الموضوع ليس بهذه السهولة بسبب النمو الكبير المتزايد، ستكون النتيجة غير منطقية وليست عملية، إذا لابد من فهم دلالي آلي آني للمحتوى وأخذ النتيجة بشكل لحظي.

2 الغاية من بناء هذا المنتج

إذا سيكون الحل هو بناء منتج يعمل على الفهم الدلالي الآني الآلي للمحتوى، هذه الخدمة التي سيتم تقديمها من خلال المشروع سيتم استغلالها في عدة نواحي من عمليات معالجة المحتوى، سواء على مستوى الفلترة، أو على مستوى التصنيف، أو على مستوى محركات البحث أيضا.

3 تحديد الجهة المهتمة بالمشروع

1- شركة ترغب بحجب محتوى ما عن موظفيها أو مستخدم الحواسيب ضمن الشركة وضمن الشبكة الخاصة بالشركة، من خلال Proxy server يتم من خلال الحجب وعدم الحجب بناء على محتوى الصفحة وبناء على القواعد الموضوعة من قبل مديري هذه الشركة "أي للاستخدام الجماعي"؛

2- شخص يرغب بحجب محتوى ما عن أولاده أو مدرس عن طلابه وغير ذلك من الأشخاص الذين يرغبون بالحجب، من خلال إضافة يتم إضافتها على المتصفح chrome extention "الاستخدامات الفردية"؛

3- جهة ترغب بتصنيف المحتوى الموجود لديها وتوزيعه على مجالات محددة وذلك من خلال تطبيق موبايل يعمل على تصنيف الأخبار الواردة حسب المجال الذي تندرج تحته.

4 أجزاء المشروع الأساسية

يتألف المشروع كمنتج نهائي من خدمة ويبweb service تعمل على الفهم الدلالي لمتحوى، تم الاستفادة من هذه الخدمة من خلال ثلاثة تطبيقات:

1- فلترة صفحات الإنترنت من خلال هذا المخدم الذي يخدم كل الموجودين على الشبكة ضمن شركة مثلا Proxy server؛

2- إضافة على المتصفح تعمل على فلترة صفحات الإنترنت المطلوبة من خلال المتصفح Chrome extention؛

3- تطبيق موبايل يهدف إلى تصنيف الأخبار الواردة من قناة ما.

الفصل الثامن

حالات استخدام النظام

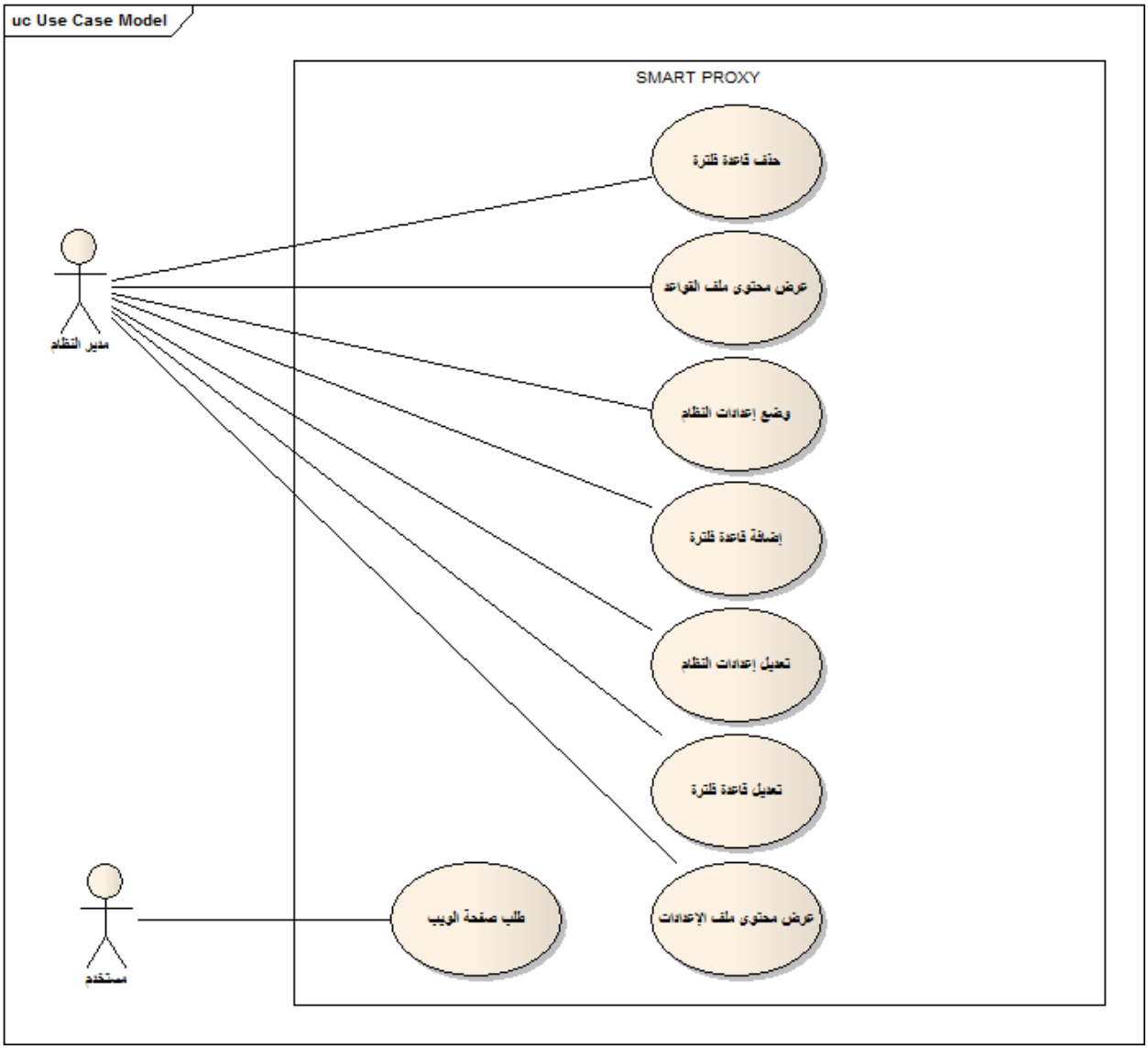
1 مقدمة

مخطط حالات الاستخدام يظهر نقاط التفاعل بين النظام وبين مستخدمي النظام، وهي تعطي فكرة أولية عن آلية عمل النظام من وجهة نظر المستخدم، فلا يظهر من النظام إلا العمليات التي يتفاعل فيها مع محيطه الخارجي، أما الآليات الداخلية لا تظهر ضمن هذه المخططات.

2 مخطط حالات الاستخدام الأولي

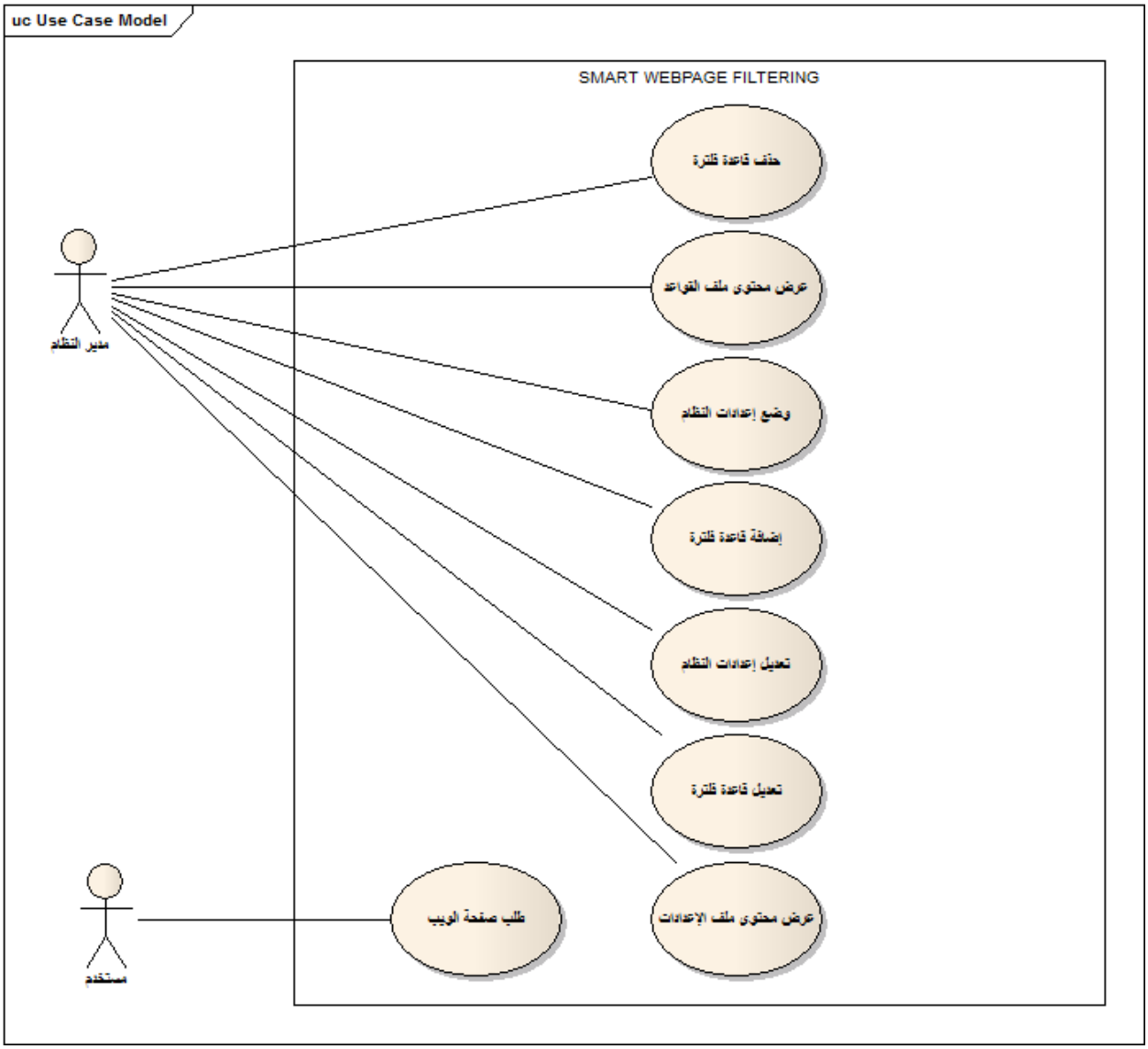
المخطط العام لحالات الاستخدام، يشمل حالات الاستخدام العامة دون الخوض في تفاصيلها الدقيقة، يعطي تصورا مبدئيا عم مهمات النظام ومن هم الفاعلون المشاركون فيها.

2.1 مخطط حالات استخدام مخدم البروكسي Smaz Proxy



الشكل 44 مخطط حالات الاستخدام لمخدم البروكسي

2.2 مخطط حالات استخدام مصنف الأخبار Smaz Filter



الشكل 45 مخطط حالات الاستخدام لقارئ الأخبار

3 حالات استخدام مخدم البروكسي Smaz Proxy

3.1 وضع إعدادات النظام

حالة الاستخدام	تعيين إعدادات النظام
رقم حالة الاستخدام	1
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	تعيين إعدادات النظام والمعاملات اللازمة للعمل.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام وضع الإعدادات الخاصة بالنظام؛ 2- يقوم مدير النظام بتحديد الإعدادات من خلال وضع القيم اللازمة لكل إعداد ضابط للنظام: – عنوان خدمة الويب؛ – البحث ضمن dbpedia؛ – خوارزمية فك الغموض المتبعة؛ – رد الضمائر إلى أصلها أم لا.
الشروط اللاحقة	تم وضع إعدادات النظام من قبل مدير النظام.
الطرق البديلة	لايوجد.

جدول 5 وضع إعدادات النظام

3.2 تعديل إعدادات النظام

حالة الاستخدام	تعديل إعدادات النظام
رقم حالة الاستخدام	2
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	تعديل إعدادات النظام والمعاملات اللازمة للعمل.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	3- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام تعديل الإعدادات الخاصة بالنظام؛ 4- يقوم مدير النظام بتعديل الإعدادات التي يريد من خلال وضع القيم اللازمة لكل إعداد ضابط للنظام: – عنوان خدمة الويب؛ – إضافة الأوزان إلى الآباء؛ – البحث ضمن dbpedia؛ – خوارزمية فك الغموض المتبعة؛ – رد الضمائر إلى أصلها أم لا.
الشروط اللاحقة	تم تعديل إعدادات النظام من قبل مدير النظام.
الطرق البديلة	لايوجد.

جدول 6 تعديل إعدادات النظام

3.3 إضافة قاعدة فلترة

حالة الاستخدام	إضافة قاعدة فلترة.
رقم حالة الاستخدام	3
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	إضافة قاعدة إلى مجموعة القواعد التي سيتم على أساسها أخذ قرار الحجب أم لا.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام إضافة قاعدة فلترة إلى مجموعة القواعد المتواجدة؛ 2- يقوم مدير النظام بوضع كل مجال ونسبة وروده ضمن الصفحة؛ 3- وضع الكلمات التي يريد ورودها ضمن الصفحة في حال تحقق الشرط الأول؛ 4- وضع مرادفات الكلمات التي يريد ورودها ضمن الصفحة؛ 5- وضع الحدث اللازم حدوثه عن تحقق القاعدة وهو حجب الصفحة أم لا.
الشروط اللاحقة	تم إضافة قاعدة فلترة إلى مجموعة القواعد المتواجدة مسبقا.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 7 إضافة قاعدة فلترة

3.4 تعديل قاعدة فلترة

حالة الاستخدام	تعديل قاعدة فلترة.
رقم حالة الاستخدام	4
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	تعديل قاعدة من قواعد الفلترة الموجودة مسبقا.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام تعديل قاعدة فلترة ضمن مجموعة القواعد المتواجدة؛ 2- في حال أراد تعديل المجالات يقوم مدير النظام بتعديل المجالات ونسبة ورودها ضمن الصفحة؛ 3- في حال أراد تعديل الكلمات يقوم المدير بتعديل الكلمات التي يريد ورودها ضمن الصفحة في حال تحقق الشرط الأول؛ 4- في حال أراد تعديل المرادفات يقوم بتعديل مرادفات الكلمات التي يريد ورودها ضمن الصفحة؛ 5- في حال أراد تعديل حدث القاعدة يقوم المدير بتعديل الحدث اللازم حدوثه عن تحقق القاعدة وهو حجب الصفحة أم لا.
الشروط اللاحقة	تم تعديل قاعدة فلترة موجودة ضمن مجموعة القواعد المضافة مسبقا.

الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول8 تعديل قاعدة فلترة

3.5 طلب صفحة ويب

حالة الاستخدام	طلب صفحة ويب
رقم حالة الاستخدام	5
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	طلب صفحة ويب ومعالجتها.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام طلب صفحة ويب؛ 2- يقوم المستخدم بطلب الصفحة التي يريد؛ 3- يمر هذا الطلب إلى squid proxy server؛ 4- بعدها يتم تمرير الطلب إلى icap server ليقوم بمعالجة الصفحة المررة دلاليا، 5- يتم الحصول على نسبة كل مجال ضمن الصفحة ومقارنة النتيجة بالقواعد الموجودة والإقرار فيما اذا كانت الصفحة ستحجب أم لا؛ 6- في حال كان القرار هو حجب الصفحة يتم رجوع النتيجة بعكس مسار الطلب

	وعرض صفحة الحجب للمستخدم؛ 7- في حال كان القرار هو عدم حجب الصفحة يتم يتم رجوع النتيجة بعكس مسار الطلب وعرض الصفحة للمستخدم؛
الشروط اللاحقة	تم طلب صفحة ويب ومعالجتها والإقرار فيما إذا كان سيتم عرض الصفحة أم لا.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول9 طلب صفحة ويب

3.6 حذف قاعدة فلترة

حالة الاستخدام	حذف قاعدة فلترة.
رقم حالة الاستخدام	6
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	حذف قاعدة من قواعد الفلترة الموجودة مسبقا.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام حذف قاعدة فلترة ضمن مجموعة القواعد المتواجدة؛ 2- يقوم المدير بتحديد القاعدة التي يريد حذفها؛ 3- يطلب المدير الأمر حذف القاعدة؛

	4- يقوم النظام بحذف القاعدة وحفظ التعديلات الحاصلة في حال أكد المدير أمر الحذف.
الشروط اللاحقة	تم حذف قاعدة فلترة موجودة ضمن مجموعة القواعد المضافة مسبقا.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 10 حذف قاعدة فلترة

3.7 عرض محتوى ملف القواعد

حالة الاستخدام	عرض محتوى ملف القواعد.
رقم حالة الاستخدام	7
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	عرض محتوى ملف قواعد الحجب التي تم إضافتها مسبقا.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام عرض محتوى ملف القواعد والاطلاع على القواعد الموضوعة مسبقا؛ 2- يطلب المدير من النظام عرض ملف القواعد؛ 3- يتم عرض الملف ضمن نافذة تحوي القواعد بالترتيب الموجود مسبقا مع الحدث الذي يرافق كل قاعدة.

الشروط اللاحقة	تم عرض محتوى ملف القواعد الموضوعة مسبقا من قبل مدير النظام.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 11 عرض محتوى ملف القواعد

3.8 عرض محتوى ملف الإعدادات

حالة الاستخدام	عرض محتوى ملف الإعدادات
رقم حالة الاستخدام	8
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	عرض محتوى ملف الإعدادات التي تم تحديدها مسبقا.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام عرض محتوى ملف الإعدادات والاطلاع على المحددات الموضوعة مسبقا؛ 2- يطلب المدير من النظام عرض ملف الإعدادات؛ 3- يتم عرض الملف ضمن نافذة تحوي االإعدادات الموجودة ضمن الملف والتي تظهر بشكل إعداد مع قيمة هذا الإعداد.
الشروط اللاحقة	تم عرض محتوى ملف الإعدادات الموضوعة مسبقا من قبل مدير النظام.

الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 12 عرض محتوى ملف الإعدادات

4 حالات استخدام إضافة المتصفح Smaz Filter

حالات استخدام إضافة المتصفحة هي نفسها حالات استخدام مخدم البروكسي، الفرق الوحيد في هذا التطبيق أنه موضوع من أجل الأستخدامات الفردية، وأن طلب خدمة الويب التي تزودنا بالفهم الدلالي للصفحة تتم عن طريق الإضافة مباشرة ، بينما في المخدم كانت الخدمة يتم طلبها من خلال مخدم الويب.

5 حالات استخدام مصنّف الأخبار Smaz Reader

5.1 عرض الأخبار التابعة لمجال معين

حالة الاستخدام	عرض الأخبار التابعة لمجال معين.
رقم حالة الاستخدام	1
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	عرض الأخبار التابعة لمجال معين.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام عرض الأخبار التابعة لمجال معين؛

	2- يقوم النظام بعرض قائمة المجالات المتاحة؛ 3- يختار المدير المجال الذي يريد عرض أخباره؛ 4- يقوم النظام بالحصول على أخبار هذا المجال الموجودة ضمن قاعدة معطيات خاصة؛ 5- يقوم النظام بعرض الأخبار التابعة لهذا المجال.
الشروط اللاحقة	تم عرض الأخبار التابعة لمجال معين من قبل مستخدم النظام.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول13 عرض الأخبار التابعة لمجال معين

5.2 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها

حالة الاستخدام	تحديد مجالات الأخبار المرغوب عرضها.
رقم حالة الاستخدام	2
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	تحديد المجالات التي يرغب المستخدم بعرض الأخبار الخاصة بها.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام تحديد المجالات التي يرغب بعرض الأخبار الخاصة بها؛

	2- يقوم النظام بعرض قائمة تحوي كل المجالات المتاحة؛ 3- يقوم مستخدم النظام بتحديد المجالات التي يريد؛ 4- يتم حفظ التعديلات في حال أكد المستخدم ذلك.
الشروط اللاحقة	تم تحديد المجالات التي يرغب المستخدم بعرض أخبارها.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول14 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها

5.3 وضع إعدادات التطبيق

حالة الاستخدام	وضع إعدادات التطبيق
رقم حالة الاستخدام	3
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	وضع إعدادات تطبيق الموبايل.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام وضع إعدادات تطبيق الموبايل الخاص بتصنيف الأخبار؛ 2- يقوم المستخدم بتحديد درجة حساسية انتماء الخبر لمجال معين؛ 3- يقوم المستخدم بتحديد هل يريد إظهار عدد الأخبار ضمن كل مجال من

	المجالات أم لا؛ 4- يطلب المستخدم من النظام حفظ الإعدادات.
الشروط اللاحقة	تم وضع إعدادات التطبيق من قبل مستخدم النظام.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 15 وضع إعدادات التطبيق

5.4 عرض المجالات المتواجدة

حالة الاستخدام	عرض المجالات المتواجدة.
رقم حالة الاستخدام	4
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	عرض المجالات المتواجدة والتي تحوي مجموعة من الأخبار.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام معرفة المجالات الموجودة ضمن التطبيق؛ 2- يقوم النظام بعرض المجالات التي سبق وقد تم تحديدها من قبل المستخدم من بين كل المجالات المتوافرة.

الشروط اللاحقة	تم عرض المجالات المرغوب فيها.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 16 عرض المجالات المتواجدة

5.5 إضافة قناة أخبار

حالة الاستخدام	إضافة قناة أخبار.
رقم حالة الاستخدام	5
المثلين الأوليين	مستخدم النظام
المثلين الثانويين	لايوجد.
توصيف مختصر	إضافة قناة أخبار ليتم تحصيل الأخبار منها.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام إضافة قناة أخبار إلى قائمة القنوات المتواجدة؛ 2- يطلب المستخدم الأمر إضافة قناة؛ 3- يقوم المستخدم بإدخال عنوان القناة؛ 4- يقوم المستخدم بإدخال رابط صفحة الأخبار؛ 5- يقوم النظام عندها بتحصيل الأخبار الموجودة ضمن الصفحة ومعالجتها وتحديد المجالات التي ينتمي إليها كل خبر من الأخبار.

الشروط اللاحقة	تم إضافة قناة أخبار إلى قائمة القنوات الموجودة مسبقا.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 17 إضافة قناة أخبار

5.6 عرض القنوات المتواجدة

حالة الاستخدام	عرض قنوات الأخبار الموجودة مسبقا.
رقم حالة الاستخدام	6
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	عرض قنوات الأخبار التي سبق وتم إضافتها ومعالجة الأخبار ضمنها.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام عرض القنوات التي تم إضافتها مسبقا إلى التطبيق؛ 2- يقوم النظام بالحصول على القنوات المتواجدة لديه ضمن قاعدة معطياته؛ 3- يقوم النظام بعرض القنوات الموجودة للمستخدم.
الشروط اللاحقة	تم عرض القنوات المتواجدة مسبقا لمستخدم النظام.
الطرق البديلة	لايوجد.

لاستثناءات	لايوجد.

جدول 18 عرض القنوات المتواجدة

5.7 حذف قناة

حالة الاستخدام	حذف قناة أخبار موجودة مسبقا.
رقم حالة الاستخدام	7
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	حذف قناة أخبار موجودة مسبقا.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام حذف قناة أخبار سبق وتم إضافتها؛ 2- يقوم النظام بعرض القنوات المتواجدة ضمن التطبيق؛ 3- يحدد المستخدم القناة التي يرغب بحذفها؛ 4- يقوم النظام بحذف القناة وبالتالي حذف كل الأخبار الموجودة ضمن هذه القناة.
الشروط اللاحقة	تم حذف قناة أخبار من قبل مستخدم النظام.
الطرق البديلة	لايوجد.
لاستثناءات	لايوجد.

جدول 19 حذف قناة

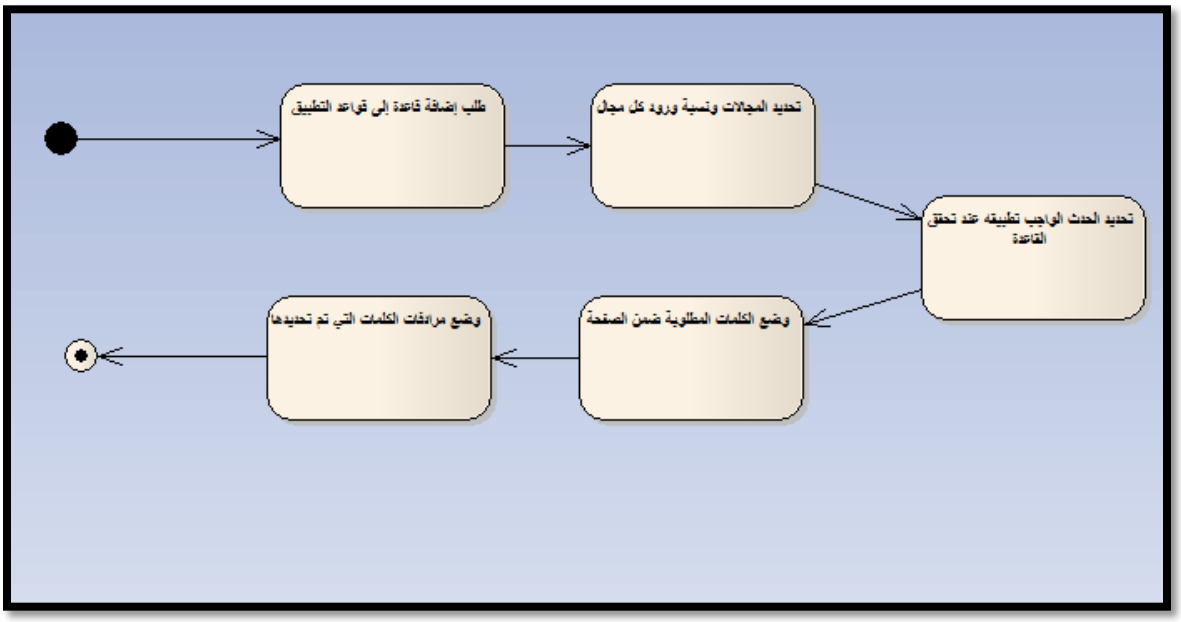
5.8 وضع إعدادات النظام

حالة الاستخدام	تعيين إعدادات النظام
رقم حالة الاستخدام	8
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لايوجد.
توصيف مختصر	تعيين إعدادات النظام والمعاملات اللازمة للعمل.
الشروط المسبقة	لايوجد.
التدفق الأساسي للأحداث	5- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام وضع الإعدادات الخاصة بالنظام؛ 6- يقوم مدير النظام بتحديد الإعدادات من خلال وضع القيم اللازمة لكل إعداد ضابط للنظام: – عنوان خدمة الويب؛ – رفع الأوزان إلى الآباء؛ – البحث ضمن dbpedia؛ – خوارزمية فك الغموض المتبعة؛ – رد الضمائر إلى أصلها أم لا.
الشروط اللاحقة	تم وضع إعدادات النظام من قبل مدير النظام.
الطرق البديلة	لايوجد.

جدول20 وضع إعدادات النظام

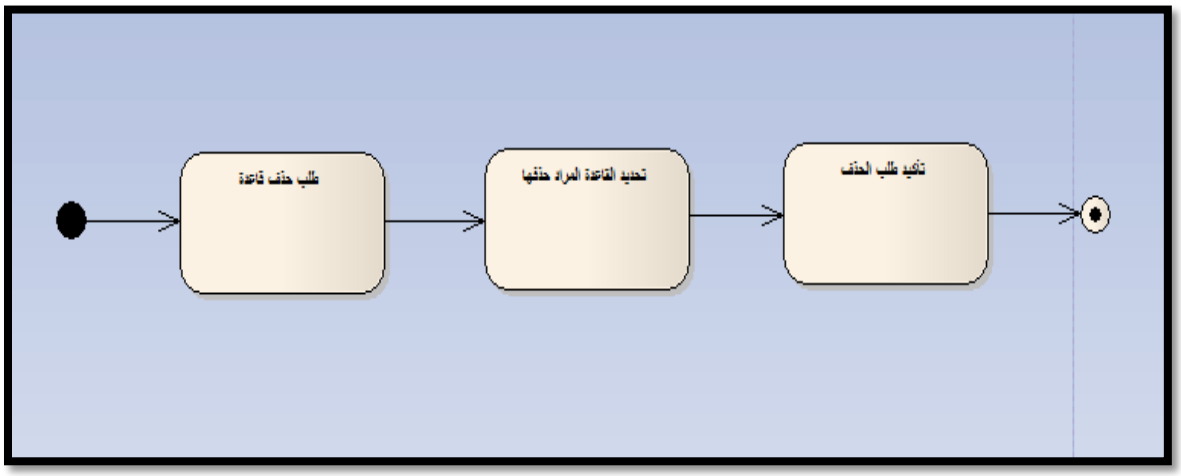
مخططات الأنشطة

إضافة قاعدة إلى قواعد الفلترة



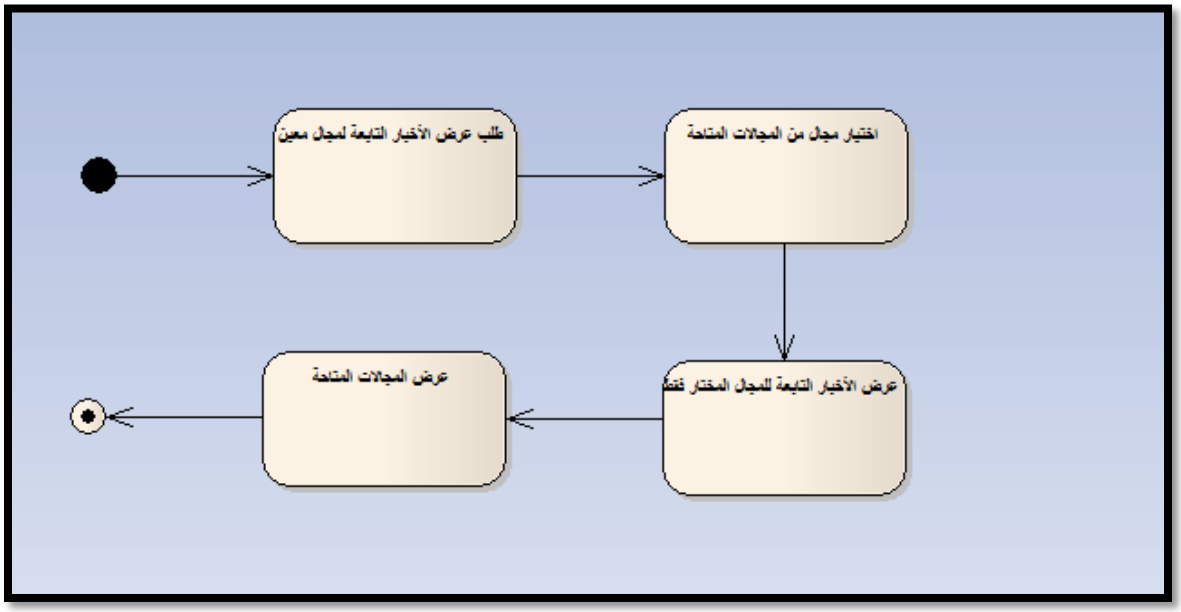
الشكل 46 إضافة قاعدة إلى قواعد الفلترة

حذف قاعدة



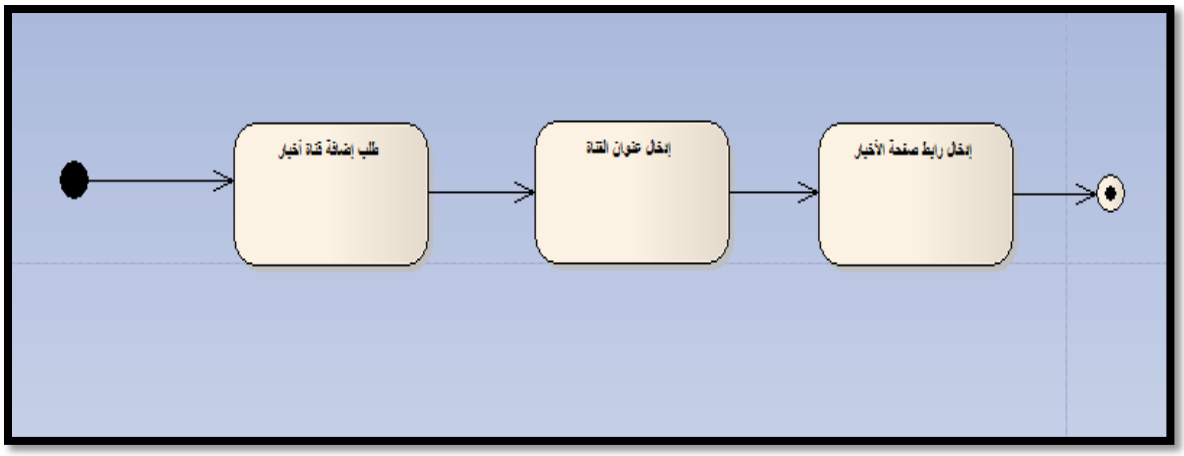
الشكل 47 حذف قاعدة

عرض الأخبار التابعة لمجال معين



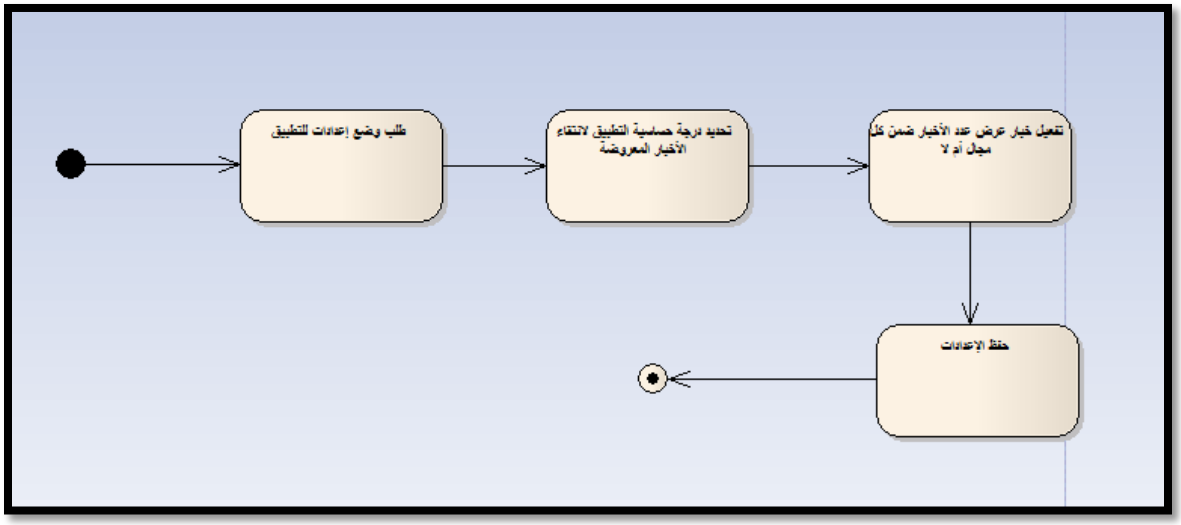
الشكل 48 عرض الأخبار التابعة لمجال معين

إضافة قناة أخبار



الشكل 49 إضافة قناة أخبار

وضع إعدادات التطبيق

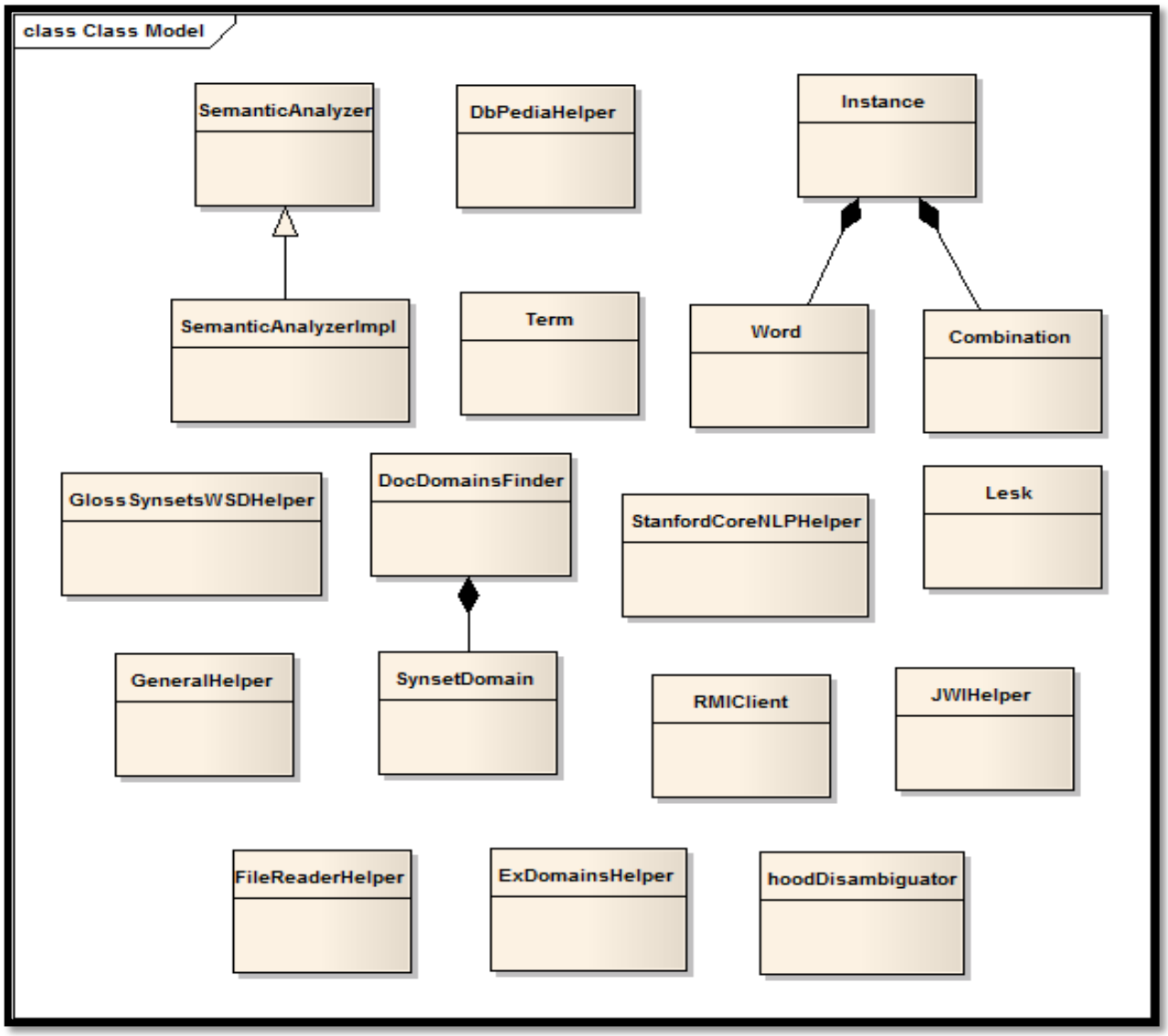


الشكل 50 وضع إعدادات للتطبيق

الفصل التاسع

مخطط الصفوف المفاهيمي

1 مخطط الصفوف المفاهيمي لمخدم البروكسي Smaz Proxy



الشكل 51 مخطط الصفوف المفاهيمي لمخدم البروكسي

2 توصيف الصفوف المفاهيمية لمخدم البروكسي

الصف	التوصيف
SemanticAnalyzer	– واجهة خدمة الويب التي تقدم خدمتين أساسيتين هما تحميل المكتبات ومعالجة طلب فهم محتوى.
SemanticAnalyzerImpl	– تحقيق الواجهة السابقة يعمل هذا الصف على إعادة تعريف الخدمات الموجودة في الصف السابق وتحقيقها؛ – يشكل هذا الصف الخدمة التي يتم الاستعانة بها من أجل فهم المحتوى.
RMIClient	– يعمل هذا الصف كزبون للخدمة السابقة والذي يقوم بطلب معالجة محتوى وفهمه من الخدمة.
Instance	– هذا الصف يقوم بإيجاد جميع التركيبات الممكنة من معاني الكلمات الموجودة في النافذة ويقوم بالمرور على كل تركيبة وحساب رصيدها، ومن ثم اختيار التركيبة ذات الرصيد الأعلى لتكون كلمات هذه التركيبة هي المعاني الصحيحة للكلمات الموجودة في النافذة

Combination	– هذه الصف يحوي مجموعة من الكلمات بحجم النافذة المحددة وبمعاني محددة ليتم فحص توافقها مع بعضها البعض لإيجاد الرصيد (score)النهائي لهذه التركيبة والذي يعبر عن مدى ارتباط كلماتها ببعضها البعض.
Word	– هذا الصف يعبر عن الكلمة حيث يتم تخزين فيه موقع هذه الكلمة في الجملة ومعاني هذه الكلمة جميعها بالإضافة إلى افضل معنى لها (سينتج معنا من الخوارزمية).
Lesk	– إزالة غموض الكلمات من خلال خوارزمية lesk.
hoodDisambiguator	– يعمل هذا الصف على فك غموض جملة من خلال الاستعانة بخوارزمية hood في إزالة غموض الكلمات.
Term	– يشكل هذا الصف مفهوم كلمة من الكلمة بحد ذاتها ومكانها في الجملة والجذر الخاص بها.
StanfordCoreNLPHelper	– يقدم هذا الصف عمليات مساعدة من أجل التعامل مع مكتبة StanfordCoreNLP

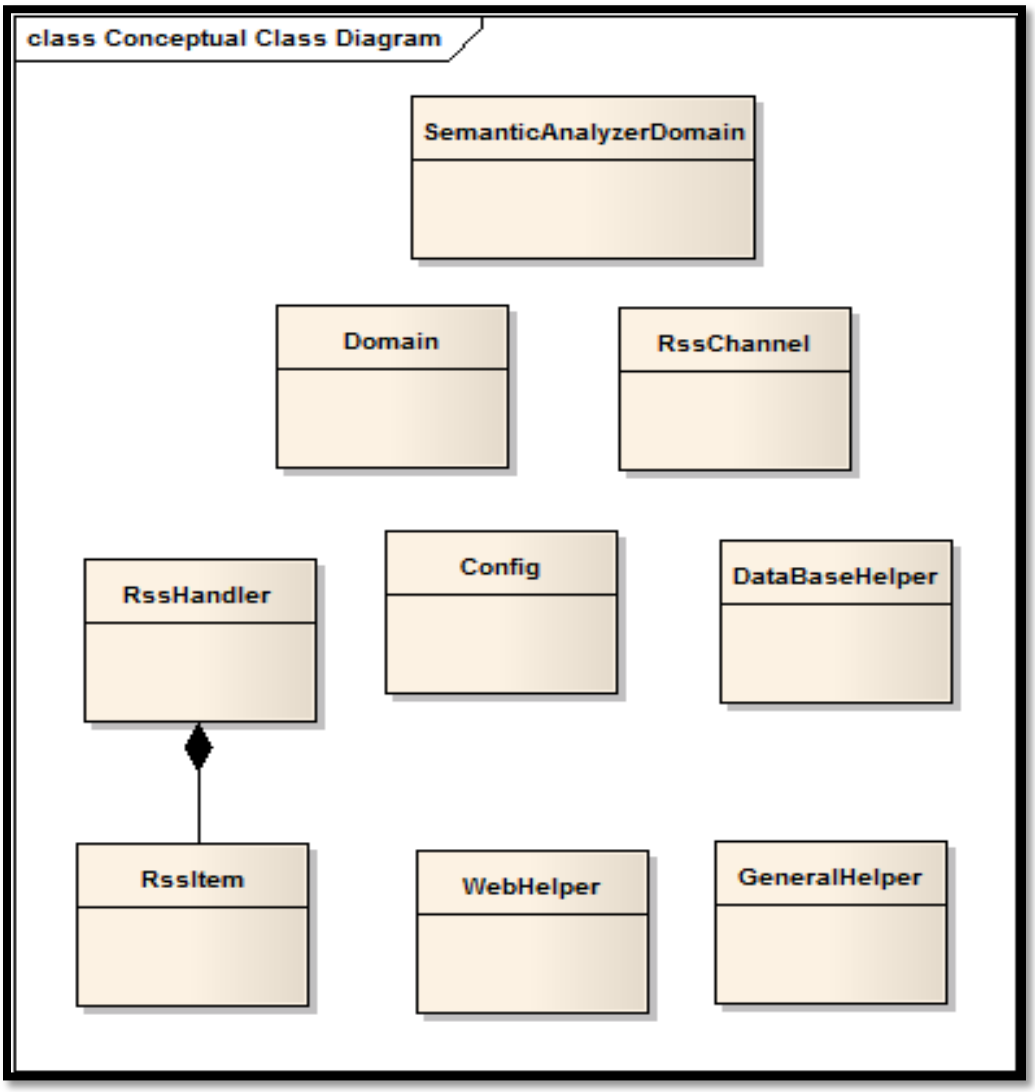
	مثل الحصول على أصل الكلمة.
JWIHelper	– يقدم هذا الصف عمليات مساعدة من أجل التعامل مع مكتبة wordnet مثل الحصول على المجموعة الأشهر لكلمة، والحصول على المجموعات التي تنتمي لها كلمة ما.
DbPediaHelper	– يقدم هذا الصف عمليات مساعدة من أجل التعامل مع أنطولوجية dbpedia مثل الاستعلام ضمنها والحصول على معنى كلمة ما.
ExDomainsHelper	– يقدم هذا الصف عمليات مساعدة من أجل التعامل مع wordnet domains مثل الحصول على آباء مجال ما والحصول على المجالات الموجودة وتحميلها مع الأوزان إلى الذاكرة.
FileReaderHelper	– يقدم هذا الصف عمليات من أجل قراءة محتوى بصيغ مختلفة سواء كان word, pdf, html .....
GlossSynsetsWSDHelper	– يقدم هذا الصف عمليات مساعدة من أجل خوارزمية إزالة الغموض المتعلقة ب gloss

	.synset
GeneralHelper	– يقدم هذا الصف عمليات مساعدة من أجل عمليات معالجة اللغات الطبيعية من تحميل الكلمات التي لا معنى لها وتقسيم الجمل والكلمات.
SynsetDomain	– يقدم هذا الصف عمليات مساعدة من أجل التعامل مع domains مثل إسناد الاسم والوزن لكل مجال ضمن مجموعة ما.
DocDomainsFinder	– يقدم هذا الصف عمليات مساعدة من أجل التعامل مع domains مثل رفع الأوزان إلى الآباء وعمل تنتظيم لقيم الأوزان ضمن مجال محدد.

جدول 21 توصيف الصفوف المفاهيمية لمخدم البروكسي

3 مخطط الصفوف المفاهيمي للقارئ Smaz Reader

- هنا تتكرر نفس الصفوف التي لها علاقة بالفهم الدلالي للمحتوى لذلك لم يتم إعادة رسمها؛
- تم وضع الصفوف التي لها علاقة بتطبيق الموبايل الخاص بتصنيف الأخبار.



الشكل52 مخطط الصفوف المفاهيمي لقارئ الأخبار

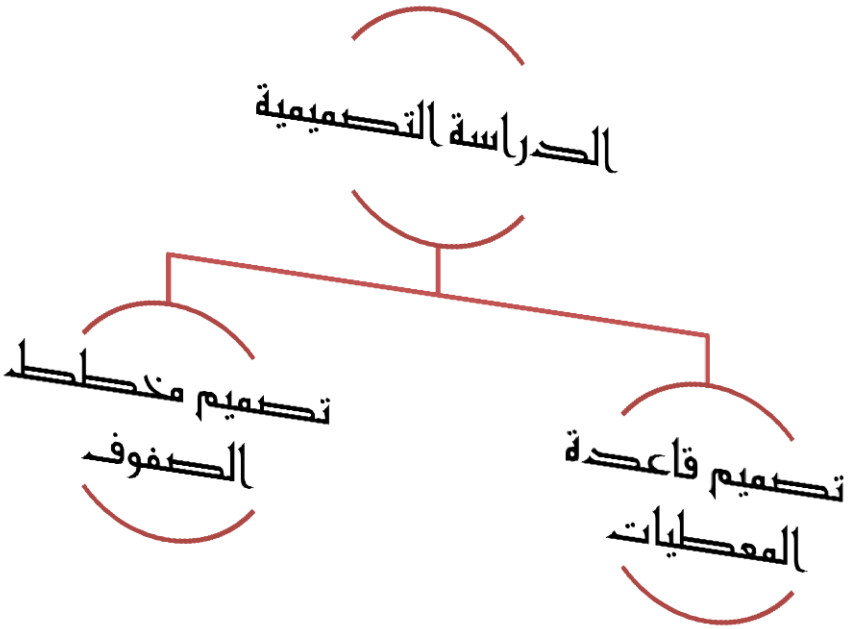
4 توصيف الصفوف المفاهيمية للقارئ

الصف	التوصيف
SemanticAnalyzerDomain	– يؤمن هذا الصف مجال معين موجود ضمن محتوى ما ونسبة وجوده ضمن المحتوى؛ – عند معالجة محتوى ما يتم الحصول على قائمة من هذا الصف.
Domain	– يمثل هذا الصف مجال معين.
RssItem	– يمثل هذا الصف خبر من الأخبار الموجودة ضمن قناة الأخبار.
RssChannel	– يمثل هذا الصف قناة الأخبار.
DataBaseHelper	– يؤمن هذا الصف مجموعة من العمليات المساعدة للتعامل مع قواعد المعطيات من إضافة وحذف وتعديل للمجالات والقنوات والأخبار.
GeneralHelper	– يؤمن هذا الصف مجموعة من العمليات المساعدة مثل تحميل المجالات وإضافتها إلى القائمة.
WebHelper	– يؤمن هذا الصف مجموعة من العمليات المساعدة مثل الاتصال بالانترنت وتحميل الأخبار الموجودة ضمن قناة ما.

جدول22 توصيف الصفوف المفاهيمية للقارئ

الباب الرابع

الدراسة التصميمية

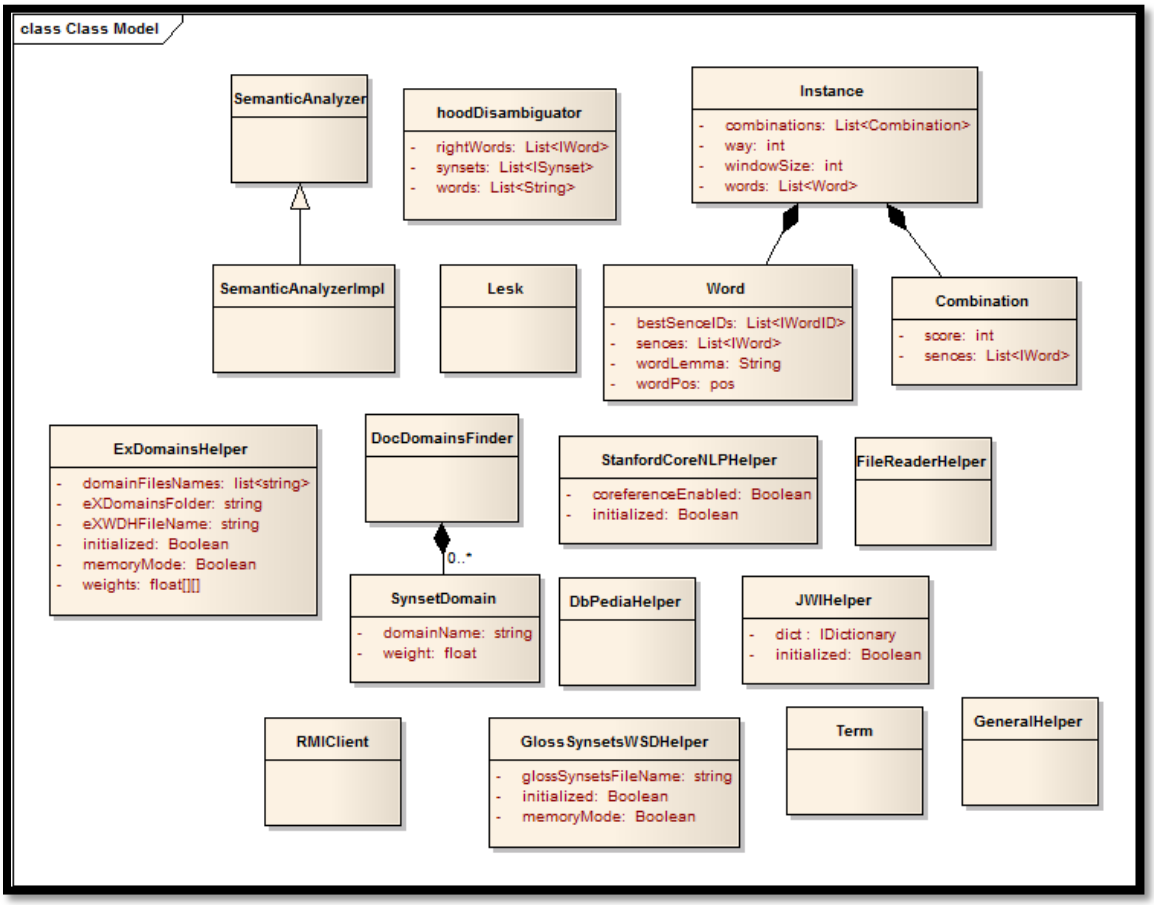


الشكل53 النقاط الأساسية ضمن مرحلة التصميم

الفصل العاشر

مخطط الصفوف

1 مخطط صفوف مخدم البروكسي Smaz Proxy



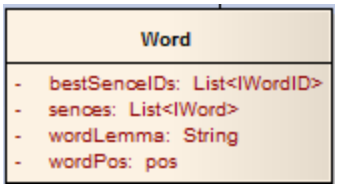
الشكل 54 مخطط صفوف مخدم البروكسي

2 شرح بعض صفوف مخدم البروكسي Smaz Proxy

2.1 hoodDisambiguator

- يستخدم هذا الصف من أجل إزالة غموض جملة عن طريق خوارزمية إزالة الغموض hood؛
- يحوي الخصائص التالية:
  - سلسلة الكلمات المقابلة بعد إزالة غموض الكلمات؛
  - سلسلة المجموعات التي سنحصل عليها عند انتهاء الخوارزمية من عملها، حيث ترد الخوارزمية المجموعة التي تنتمي إليها كل كلمة من الكلمات التي نرغب بإزالة غموضها؛
  - سلسلة من الكلمات التي نرغب بإزالة غموضها.
- يحوي العمليات التالية:
  - الحصول على الكلمات من الجملة الممررة؛
  - إزالة غموض الكلمات؛
  - إيجاد الجذر HoodRoot؛
  - إيجاد كل الآباء التابعين لمجموعة ما؛
  - الحصول على أكثر مجموعة مشهورة لكلمة ما.

2.2 Word

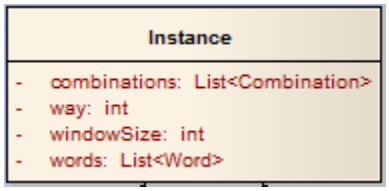


- هذا الصف يعبر عن الكلمة حيث يتم تخزين فيه موقع هذه الكلمة في الجملة ومعاني هذه الكلمة جميعها بالإضافة إلى افضل معنى لها (سينتج معنا من الخوارزمية).

2.3 Combination

- هذه الصف يحوي مجموعة من الكلمات بحجم النافذة المحددة وبمعاني محددة، ليتم فحص تواف قها مع بعضها البعض لإيجاد الرصيد (score)النهائي لهذه التركيبة والذي يعبر عن مدى ارتباط كلماتها ببعضها البعض.
- تتم عملية حساب الرصيد (score) من خلال التابع calculate_Score الذي يقوم بتشكيل جميع الأزواج الممكنة من الكلمات الموجودة في التركيبة وايجاد التشابهات بين معاني كلمات الزوج الواحد وحساب عدد التشابهات مع الأخذ بعين الاعتبار العلاقات المرتبطة بكل كلمة والطريقة المحددة لحساب الأزواج (Heterogeneous, Homogeneous).

2.4 Instance



- هذا الصف يقوم بإيجاد جميع التركيبات الممكنة من معاني الكلمات الموجودة في النافذة ويقوم بالمرور على كل تر كبية وحساب رصيدها، ومن ثم اختيار التركيبة ذات الرصيد الأعلى لتكون كلمات هذه التركيبة هي المعاني الصحيحة للكلمات الموجودة في النافذة.
- يحوي العمليات التالية:

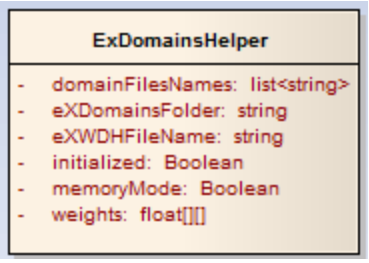
- الحصول على كل التراكيب الممكنة؛
- الحصول على أفضل تركيبة.

2.5 DbPediaHelper

- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع أنطولوجية dbpedia مثل الاستعلام ضمنها والحصول على معنى كلمة ما.
- يحوي العمليات التالية:
  - طلب خدمة من spotlight والحصول على نتيجة منها؛
  - الحصول على روابط كلمات النص المدخل ضمن dbpedia؛
  - استعلام sparql ضمن dbpedia.

2.6 ExDomainsHelper

- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع wordnet domains.
- يحوي العمليات التالية:
  - تحميل الأوزان إلى الذاكرة؛
  - الحصول على آباء مجال ما؛
  - الحصول على المجالات المتواجدة.



2.7 FileReaderHelper

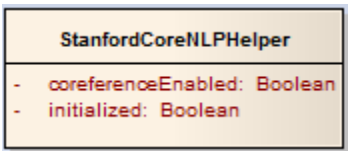
- يقدم هذا الصف عمليات من أجل قراءة محتوى بصيغ مختلفة سواء كان word, pdf, html ....؛
- يقدم هذا الصف العمليات التالية:
- قراءة ملف pdf؛
- قراءة ملف DOCX؛
- قراءة ملف html؛
- الحصول على نص من ملف html؛
- الحصول على محتوى تاغات أخرى موجودة ضمن صفحة html.

2.8 GeneralHelper

- يقدم هذا الصف عمليات مساعدة من أجل عمليات معالجة اللغات الطبيعية من تحميل الكلمات التي لا معنى لها وتقسيم الجمل والكلمات.
- بعض العمليات التي يقدمها:
  - فصل الكلمات المترابطة؛
  - إزالة المحارف الغريبة؛
  - تحميل الكلمات الغير حاملة للمعنى؛
  - الحصول على الكلمات الموجودة ضمن الجمل.

StanfordCoreNLPHelper 2.9

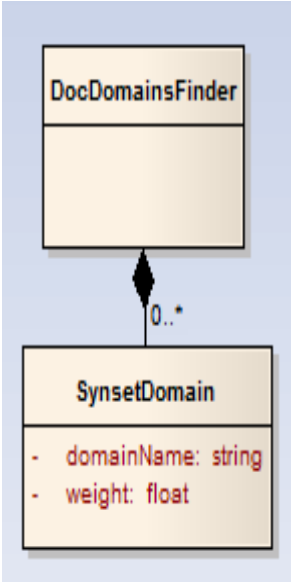
- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع مكتبة StanfordCoreNLP مثل الحصول على أصل الكلمة.



- بعض العمليات التي يقدمها:
  - الحصول على أصل الكلمة؛
  - الحصول على الكلمات الموجودة ضمن مواقع مقبولة ومحددة مسبقا.

DocDomainsFinder 2.10

- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع domains مثل رفع الأوزان إلى الآباء وعمل تنظيم لقيم الأوزان ضمن مجال محدد.
- بعض العمليات التي يقدمها:
  - إضافة وزن الكلمة إلى مجموع المجال الكلي؛
  - إضافة الوزن إلى الآباء؛
  - عمل تنظيم للأوزان ضمن مجال محدد؛
  - الحصول على المجالات والأوزان الخاصة بها من أجل محتوى ما.



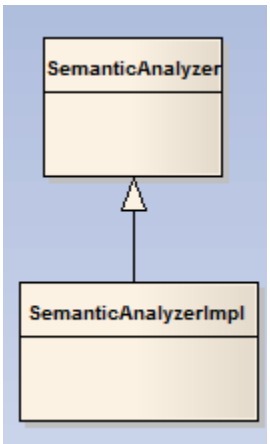
2.11 SynsetDomain

- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع domains مثل إسناد الاسم والوزن لكل مجال ضمن مجموعة ما.

2.12 JWIHelper

- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع مكتبة wordnet؛
- بعض العمليات التي يقدمها:
  - مثل الحصول على المجموعة الأشهر لكلمة؛
  - الحصول على المجموعات التي تنتمي لها كلمة ما

2.13 SemanticAnalyzer

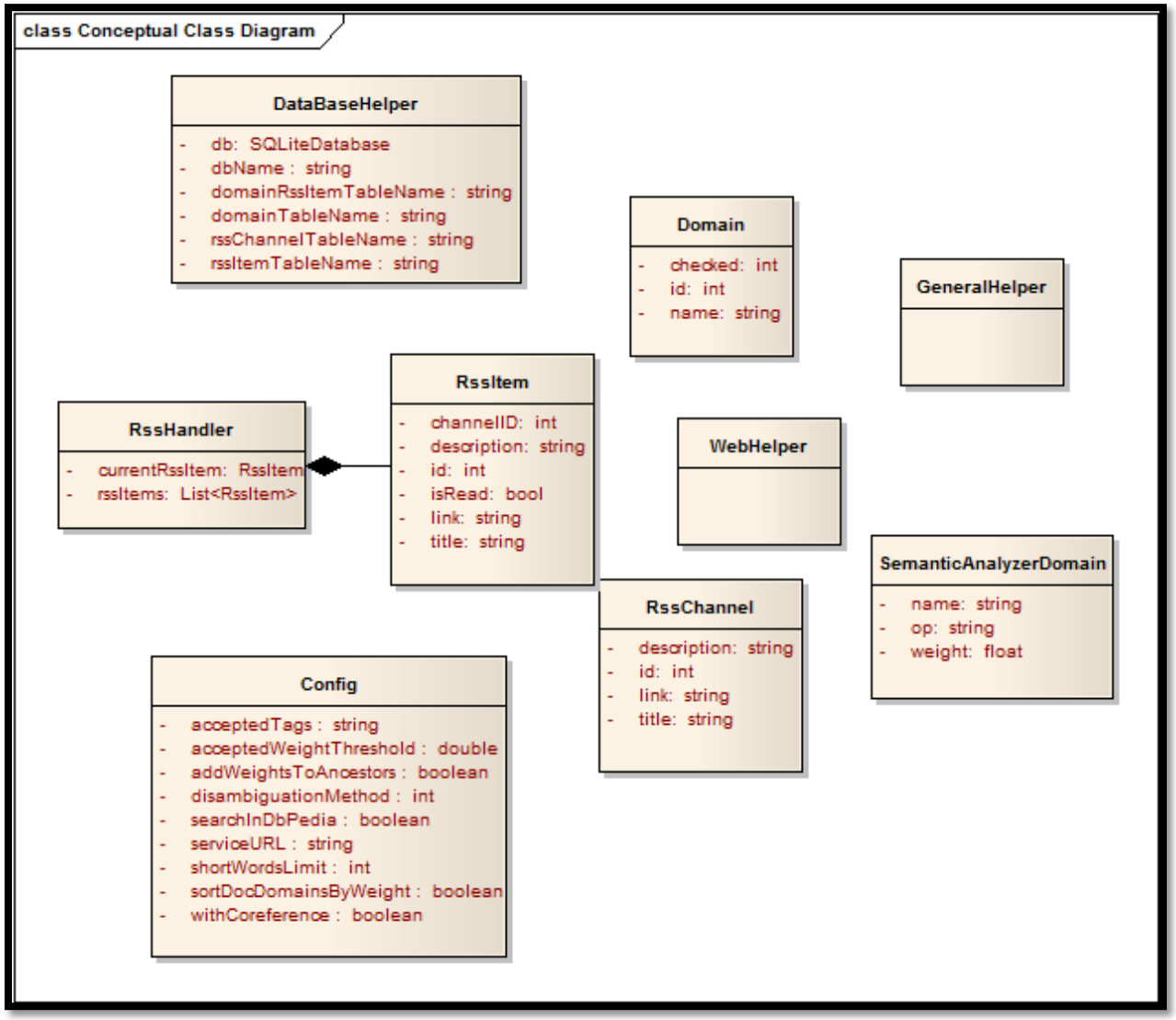


- واجهة خدمة الويب؛
- تقدم العمليتين الأساسيتين التاليتين:
  - تحميل المكتبات إلى الذاكرة؛
  - معالجة طلب فهم محتوى.

2.14 SemanticAnalyzerImpl

- تحقيق الواجهة السابقة يعمل هذا الصف على إعادة تعريف الخدمات الموجودة في الصف السابق وتحقيقها؛
- يشكل هذا الصف الخدمة التي يتم الاستعانة بها من أجل فهم المحتوى.
-

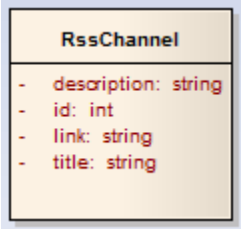
3 مخطط صفوف القارئ Smaz Reader



الشكل 55 مخطط صفوف قارئ الأخبار

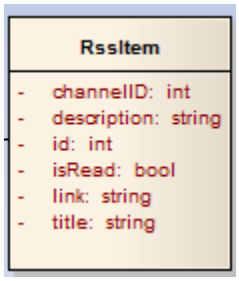
4 شرح بعض صفوف القارئ Smaz Reader

4.1 RssChannel



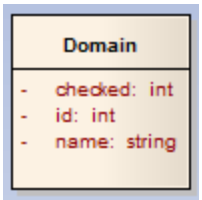
- يمثل هذا الصف قناة أخبار؛
- يحوي هذا الصف الخصائص التالية:
  - توصيف قناة الأخبار؛
  - المعرف المميز لقناة الأخبار؛
  - رابط وجود صفحة الأخبار؛
  - عنوان قناة الأخبار.

4.2 RssItem



- يمثل هذا الصف خبر من الأخبار الموجودة ضمن قناة الأخبار؛
- يحوي هذا الصف الخصائص التالية:
  - توصيف الخبر؛
  - المعرف المميز للخبر؛
  - رابط وجود الخبر؛
  - قناة الأخبار التي يتبع لها الخبر؛
  - هل تم قراءة الخبر من قبل أم لا؛
  - عنوان الخبر.

4.3 Domain



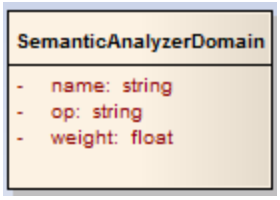
- يمثل هذا الصف مجال معين؛
- يحوي الخصائص التالية:
  - هل تم تحديد المجال من قبل أم لا؛
  - المعرف المميز للمجال؛
  - اسم المجال.

4.4 DataBaseHelper

- يؤمن هذا الصف مجموعة من العمليات المساعدة للتعامل مع قواعد المعطيات؛
- يحوي العمليات التالية:
  - إضافة مجال؛
  - إضافة كل المجالات؛
  - الحصول على كل المجالات؛
  - الحصول فقط على المجالات التي تم تحديدها؛
  - تعديل مجال ما؛
  - إضافة قناة أخبار؛
  - تعديل قناة أخبار؛
  - حذف قناة أخبار؛
  - الحصول على كل قنوات الأخبار المتواجدة؛
  - إضافة خبر؛

- تحديد المجال الخاص بخبر ما؛
- الحصول على كل الأخبار؛
- الحصول على الأخبار التابعة لمجال معين.

4.5 SemanticAnalyzerDomain

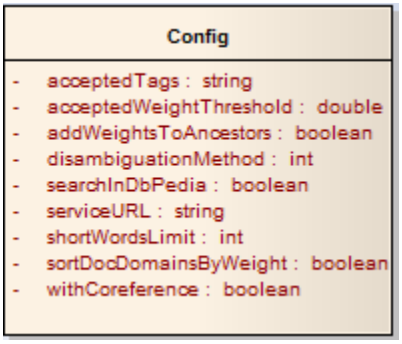


- يؤمن هذا الصف مجال معين موجود ضمن محتوى ما ونسبة وجوده ضمن المحتوى؛
- عند معالجة محتوى ما يتم الحصول على قائمة من هذا الصف؛
- يحوي الخصائص التالية:
  - اسم المجال؛
  - العملية المطبقة على الوزن الخاص بالمجال؛
  - قيمة الوزن.

4.6 WebHelper

- يؤمن هذا الصف مجموعة من العمليات المساعدة مثل الاتصال بالانترنت وتحميل الأخبار الموجودة ضمن قناة ما.
- بعض العمليات التي يقدمها هذا الصف:
  - تحميل كل الأخبار الموجودة ضمن كل القنوات المطلوبة؛
  - معالجة كل خبر من الأخبار ومعرفة المجالات التي ينتمي إليها.

4.7 Config



- يمثل هذا الصف مجموعة المتحولات التي يتم ضبط قيمها عند بداية عمل النظام؛
- يحوي الخصائص التالية:

▪ serviceURL عنوان خدمة الويب التي سيتم الاتصال بها والاستفادة من خدماتها؛

▪ sortDocDomainsByWeight متحول منطقي يأخذ قيمة true عندما نريد ترتيب domains الصفحة تنازليا حسب نسبة وردوهم ضمن الصفحة؛

▪ addWeightsToAncestors متحول منطقي يأخذ قيمة true عندما نريد رفع وزن الكلمة التابعة ل domain ما إلى آباء هذا ال domain فعادة عند ورود كلمة تنتمي إلى domain ما بنسبة 0.5 فنعطي حرية رفع هذا الوزن إلى آباء هذا ال domain أم لا من خلال هذا المتحول المنطقي؛

▪ shortWordsLimit متحول صحيح يدل على الطول الأصغري للكلمة التي يجب أخذها في الحسبان فإذا كانت قيمته مثلا 3 فعندها كل كلمة طولها أصغر من 3 لن يتم أخذها بعين الاعتبار عند المعالجة؛

▪ searchInDbPedia متحول منطقي يأخذ قيمة true عندما نريد البحث عن كلمة غير موجودة ضمن ال wordnet ضمن ال dbpedia ففي حال لم تكن الكلمة موجودة ضمن ال wordnet وكان هذا المتحول المنطقي true عندها سيتم البحث عن هذه الكلمة ضمن ال dbpedia؛

▪ dbPediaResultsLimit متحول صحيح يدل على عدد القيم التي سيتم أخذها بعين الاعتبار من أجل كل علاقة سيتم الاستفادة منها من ال dbpedia، بمعنى أنه عندما لا نجد كلمة ضمن ال wordnet سيتم البحث عنها ضمن ال dbpedia، صفحة ال dbpedia تحوي مجموعة من العلاقات التي من الممكن الاستفادة منها في محاولة لفهم معنى الكلمة، من هذه العلاقات هي علاقة subject، فمن أجل هذه

العلاقة ومن أجل الكلمة المطلوبة هناك مجموعة من القيم التي تشرح معنى هذه الكلمة وبالتالي لابد من تحديد العدد الأعظمي للقيم التي سيتم أخذها بعين الاعتبار؛

- disambiguationMethod متحول صحيح يدل على الخوارزمية المتبعة لإزالة الغموض فكما ذكرنا سابقا كنا قد حققنا مجموعة من الخوارزميات لإزالة غموض كلمة كل منها مختلفة في تعقيدها وسرعتها ودقتها عن الأخرى
  - الرقم 1 يعني أن الخوارزمية المتبعة هي خوارزمية المعنى الأكثر شهرة؛
  - الرقم 2 يعني أن الخوارزمية المتبعة هي خوارزمية WordNet Semantically Tagged glosses؛
  - الرقم 3 يعني أن الخوارزمية المتبعة هي خوارزمية Hood؛
  - الرقم 4 يعني أن الخوارزمية المتبعة هي خوارزمية Lesk.
- acceptedTags للدلالة على نوع الكلمات التي نرغب بمعالجتها ضمن النص سواء كانت فعل، اسم، صفة أو ظرف؛
- withCoreference متحول منطقي يأخذ قيمة true عندما نريد إعادة الضمائر إلى أصلها أثناء معالجة النص المطلوب؛
- dbpediaPredicates لتحديد العلاقات التي نريد معالجتها والبحث فيها ضمن أنطولوجية dbpedia عندما يكون خيار searchInDbpedia مفعل؛
- leskWindowSize لتحديد حجم النافذة الأصغري الذي نريد أن نعالجه ضمن خوازرمية إزالة الغموض الرابعة lesk؛
- leskComparsionWay لتحديد طريقة المقارنة في خوارزمية lesk.

الفصل الحادي عشر

تصميم قاعدة المعطيات

1 جداول توصيف الكيانات

RssChannel		
النمط	الشرح	الواصفة
int	رقم قناة الأخبار المميز	Id
varchar(50)	عنوان القناة	Title
varchar(50)	رابط القناة	Link
varchar(50)	توصيف القناة	Description

جدول 23 توصيف الكيان rssChannel

Domain		
النمط	الشرح	الواصفة
int	رقم المجال المميز	Id
varchar(50)	اسم المجال	Name
int	تحديد المجال	Checked

جدول 24 توصيف الكيان Domain

RssItem		
النمط	الشرح	الواصفة
int	رقم الخبر المميز	Id
varchar(50)	عنوان الخبر	Title
varchar(50)	رابط الخبر	Link
varchar(50)	توصيف الخبر	Description
Int	قراءة الخبر	isRead
Int	القناة التي يتبع لها الخبر	channelID

جدول 25 توصيف الكيان RssItem

DomainRssItem		
الواصفة	الشرح	النمط
Id	الرقم المميز	int
domainID	رقم المجال الذي يتبع له الخبر	Int
RssItemID	رقم الخبر	Int
Weight	وزن الخبر بالنسبة للمجال	float

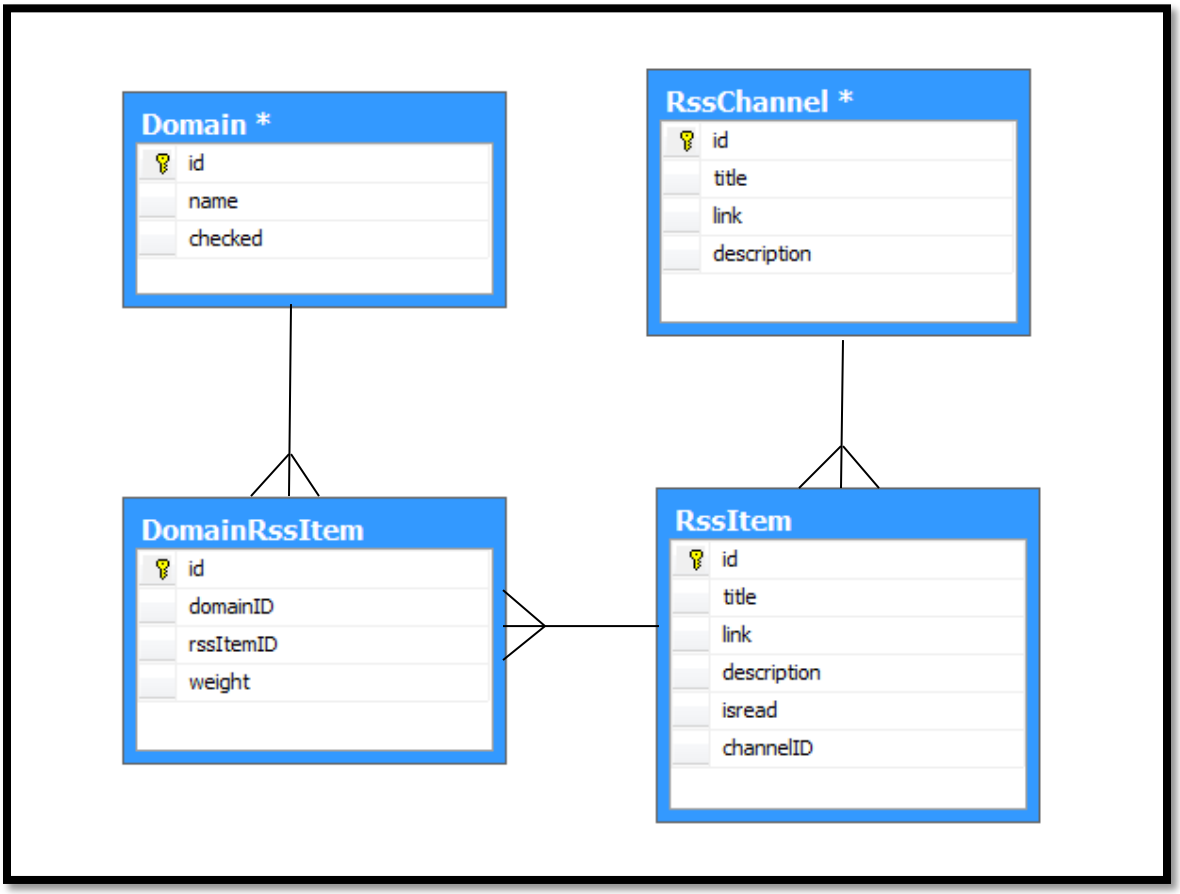
جدول 26 توصيف الكيان DomainRssItem

2 العلاقات بين الكيانات

1- هناك علاقة واحد لكثير بين الكيان RssChannel والكيان RssItem، فمن أجل كل قناة أخبار هناك مجموعة من الأخبار الموجودة ضمنها؛

2- هناك علاقة كثير لكثير بين الجدول Domain والجدول RssItem، فالمجال الواحد له أكثر من خبر والخبر والواحد يتبع لأكثر من مجال، تم كسر هذه العلاقة من خلال الجدول DomainRssItem.

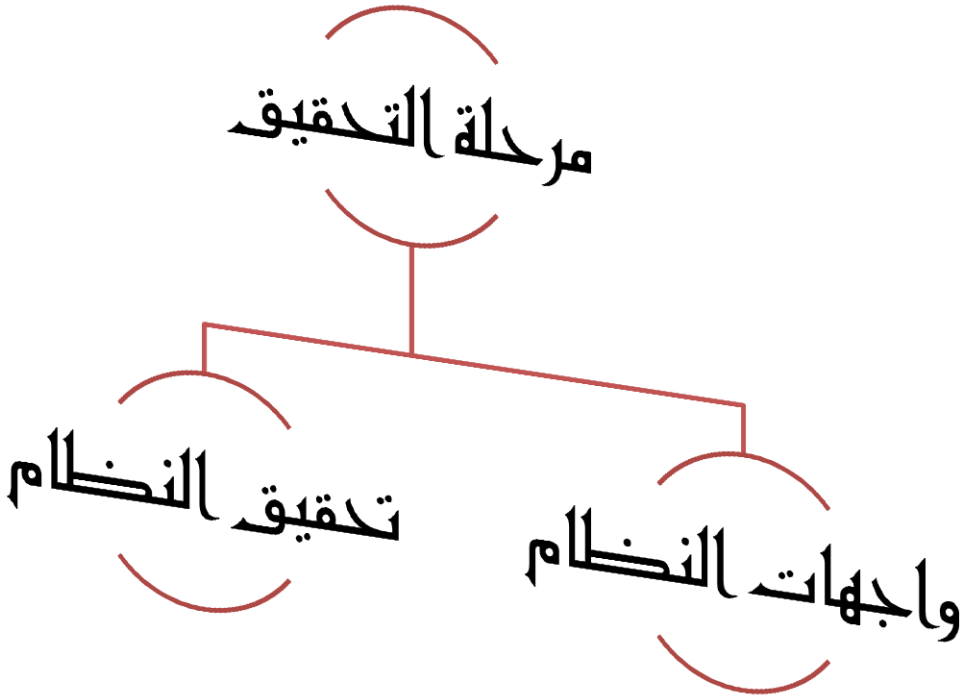
3 مخطط قاعدة البيانات



الشكل56 مخطط قاعدة المعطيات

الباب الخامس

مرحلة التحقيق

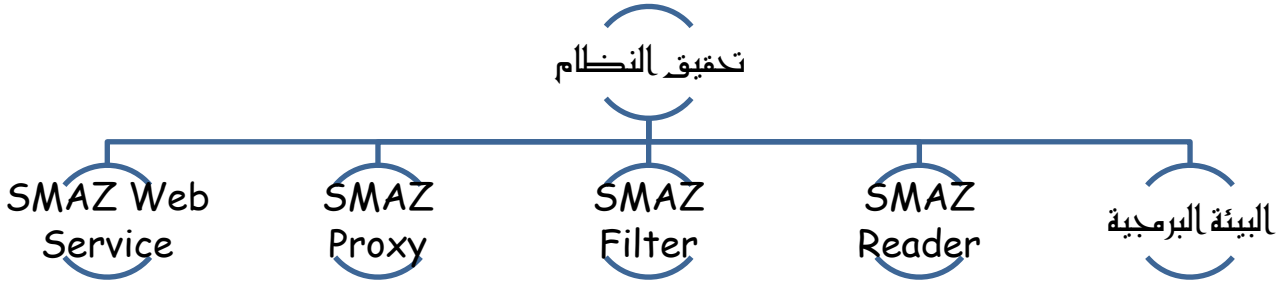


الشكل 57 النقاط الأساسية ضمن مرحلة التحقيق

الفصل الثاني عشر

تحقيق النظام

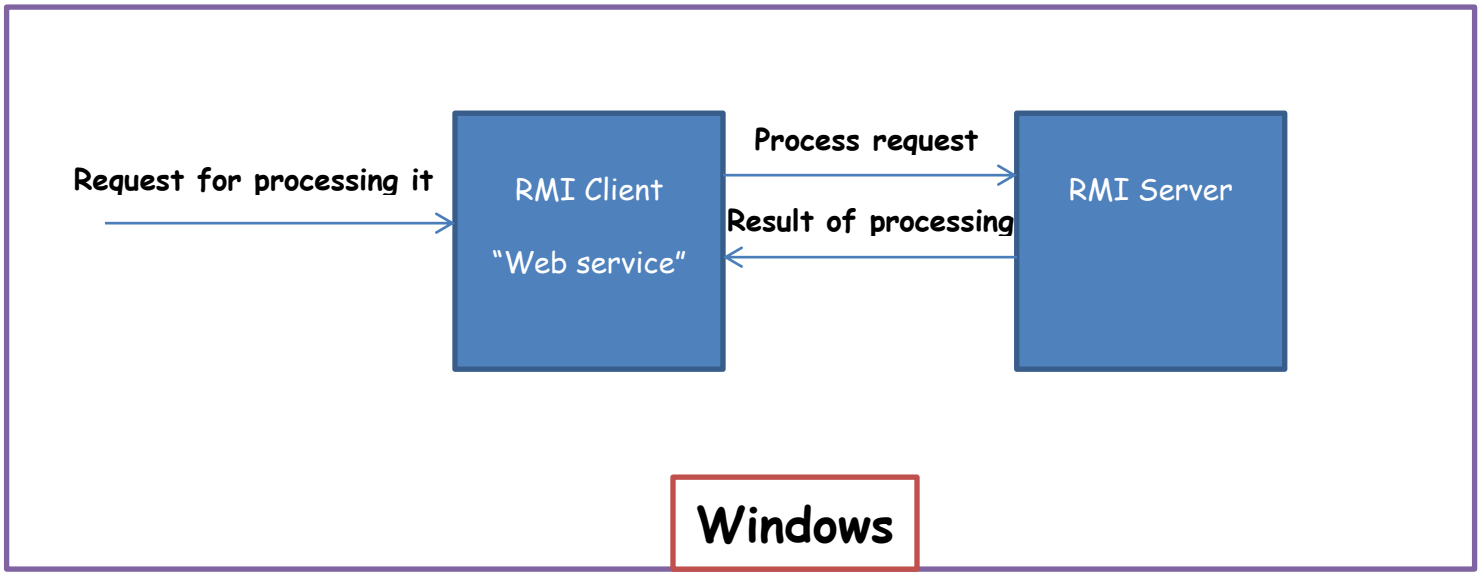
- قمنا ببناء خدمة الويب web service التي تعمل على الفهم الدلالي لمحتوى الصفحة، ومن ثم تم الاستفادة من هذه الخدمة في ثلاث تطبيقات كما ذكرنا سابقا.
  - مخدم البروكسي؛
  - القارئ ؛
  - إضافة المتصفح.
- يتم وضع إعدادات التطبيق في ملف الإعدادات؛
- من أجل التطبيقين مخدم البروكسي وإضافة المتصفح هناك قواعد يتم كتابتها ووضعها في ملف القواعد من خلال واجهة تساعد على ذلك، يتم الاستفادة من هذه القواعد للإقرار بحجب الصفحة أم لا؛
- في تطبيق الموبايل "القارئ " لا نحتاج إلى قواعد، لأن ما نقوم به هو تصنيف للأخبار وليس فلترة.



الشكل 58 النقاط الأساسية في تحقيق النظام

1 خدمة الويب Smaz web service

1.1 فكرة RMI



الشكل59 مخطط يوضح آلية عمل خدمة الويب

- Remote Method Invocation؛
- واجهتنا مشكلة أثناء العمل وهي أننا من أجل كل طلب لصفحة انترنت سيتم إعادة تنفيذ الخطوات من البداية بما فيها تحميل المكتبات التي نحتاج إليها، تحميل هذه المكتبات يأخذ وقتا طويلا وما يهمنا فعلا هو تقليل زمن تنفيذ العملية قدر الإمكان كونها تتم online؛
- من هذه الحاجة ظهرت لنا فكرة الاستفادة من تقنية RMI، حيث يكون لدينا مخدم rmi srever ويكون
- لدينا زبون rmi client، هذا المخدم مسؤول عن عمليات تحميل المكتبات والملفات اللازمة وتبقى محملة في الذاكرة طيلة الوقت، يبقى المخدم في حلقة مستمرة دائما ينتظر ورود الطلبات إليه، فمن أجل كل طلب يقوم ال icap server بطلب خدمة web service والتي تمثل بحد ذاتها ال rmi client ليقوم هذا الأخير

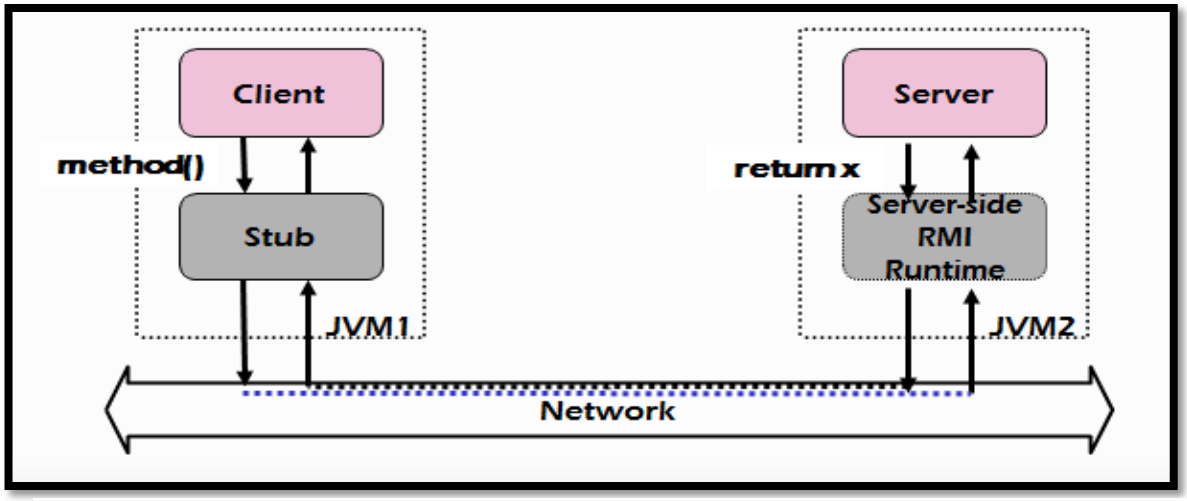
بطلب تنفيذ التابع process request الموجود عند المخدم rmi server، وتنفيذ التابع عند المخدم ورد النتيجة إلى ال rmi client ومن ثم إلى icap server؛

- لتحقيق هذه التقنية نحتاج بداية إلى remote interface هي SemanticAnalyzer extends Remote نعرف فيها التابعان:

1- load المسؤول عن تحميل المكتبات اللازمة ويرد true في حال كانت كل المكتبات قد تم تحميلها بشكل صحيح؛

2- processRequest المسؤول عن معالجة الطلبات ويرد سلسلة محرفية تحوي كل مجال domain ونسبة وروده في النص.

ونحتاج إلى Rmi server و Rmi client؛



الشكل 60 آلية عمل RMI

- ومن الجدير بالذكر أن تقنية ال RMI تخفي تفاصيل الاتصالات بين المخدم والزبون.

1.1.1 RMI Server

- نقوم ببناء rmi server والذي يقوم بعمل تحقيق للواجهة السابقة وإعادة تعريف للتوابع الموجودة فيها؛
- ضمن ال rmi server نقوم بعمل نسخة من غرض التنفيذ وتسجيله ضمن rmi registry وذلك من خلال التعليمة:

```
Registry registry = LocateRegistry.createRegistry(port);      1
SemanticAnalyzer sa = new SemanticAnalyzerImpl();      2
registry.rebind("sa", sa);      3
```

1- بناء rmi registry ليتم من خلالها تحويل غرض التنفيذ إلى غرض شبكي وحفظه ضمنها ليقوم الزبون لاحقا بالبحث عن هذا الغرض ضمنها والحصول على مؤشر عليه؛

2- عمل نسخة من غرض التنفيذ؛

3- تسجيل غرض التنفيذ ضمن rmi registry للحصول عليه لاحقا من قبل الزبون والاستفادة من عملياته التي يقدمها.

1.1.2 RMI Client

- ضمن ال rmi client يتم الحصول على نسخة من غرض التنفيذ بعد البحث عليه ضمن rmi registry واستدعاء إحدى عملياته وهي processRequest من أجل صفحة الإنترنت المطلوبة ومعالجتها عند المخدم وإرسال النتيجة النهائية للزبون

```
Registry registry = LocateRegistry.getRegistry(serverIP, port);   1

SemanticAnalyzer sa = (SemanticAnalyzer)registry.lookup("sa");   2
```

1- يتم الحصول على ال rmi registry التي تم بناؤها سابقا؛

2- يتم البحث عن غرض التنفيذ ضمن ال rmi registry والذي يحمل الاسم sa والحصول على مؤشر على غرض التنفيذ؛

3- بعد ذلك يصبح rmi client يملك مؤشر على غرض التنفيذ وبإمكانه استدعاء أي خدمة يقدمها rmi server مثل معالجة الطلبات sa.processRequest.

2 بنية ملف الإعدادات Config File

- يحوي هذا الملف مجموعة من البارامترات المحددة مسبقا ويكون موجود عند ال Icap Server وهو بدوره client لل web service المسؤولة عن حساب نسب ال domains الموجودة في صفحة الانترنت المررة أو الملف المرر للخدمة؛
- هذه البارامترات هي:
  - serviceURL عنوان خدمة الويب التي سيتم الاتصال بها والاستفادة من خدماتها؛

- getDocContentFromProxy متحول منطقي يأخذ قيمة true عندما نريد من أن يقوم بتحميل محتوى الصفحة أو الملف بنفسه وfalse عندما يقوم icap server بإرسال المحتوى إليه
- sortDocDomainsByWeight متحول منطقي يأخذ قيمة true عندما نريد ترتيب domains الصفحة تنازليا حسب نسبة وردوهم ضمن الصفحة؛
- addWeightsToAncestors متحول منطقي يأخذ قيمة true عندما نريد رفع وزن الكلمة التابعة ل domain ما إلى آباء هذا ال domain فعادة عند ورود كلمة تنتمي إلى domain ما بنسبة 0.5 فنعطي حرية رفع هذا الوزن إلى آباء هذا ال domain أم لا من خلال هذا المتحول المنطقي؛
- shortWordsLimit متحول صحيح يدل على الطول الأصغري للكلمة التي يجب أخذها في الحسبان فإذا كانت قيمته مثلا 3 فعندها كل كلمة طولها أصغر من 3 لن يتم أخذها بعين الاعتبار عند المعالجة؛
- searchInDbPedia متحول منطقي يأخذ قيمة true عندما نريد البحث عن كلمة غير موجودة ضمن ال wordnet ضمن ال dbpedia ففي حال لم تكن الكلمة موجودة ضمن ال wordnet وكان هذا المتحول المنطقي true عندها سيتم البحث عن هذه الكلمة ضمن ال dbpedia؛
- dbPediaResultsLimit متحول صحيح يدل على عدد القيم التي سيتم أخذها بعين الاعتبار من أجل كل علاقة سيتم الاستفادة منها من ال dbpedia، بمعنى أنه عندما لا نجد كلمة ضمن ال wordnet سيتم البحث عنها ضمن ال dbpedia، صفحة ال dbpedia تحوي مجموعة من العلاقات التي من الممكن الاستفادة منها في محاولة لفهم معنى الكلمة، من هذه العلاقات هي علاقة subject، فمن أجل هذه العلاقة ومن أجل الكلمة المطلوبة هناك مجموعة من القيم التي تشرح معنى هذه الكلمة وبالتالي لابد من تحديد العدد الأعظمي للقيم التي سيتم أخذها بعين الاعتبار؛

- disambiguationMethod متحول صحيح يدل على الخوارزمية المتبعة لإزالة الغموض فكما ذكرنا سابقا كنا قد حققنا مجموعة من الخوارزميات لإزالة غموض كلمة كل منها مختلفة في تعقيدها وسرعتها ودقتها عن الأخرى
  - الرقم 1 يعني أن الخوارزمية المتبعة هي خوارزمية المعنى الأكثر شهرة؛
  - الرقم 2 يعني أن الخوارزمية المتبعة هي خوارزمية WordNet Semantically Tagged glosses؛
  - الرقم 3 يعني أن الخوارزمية المتبعة هي خوارزمية Hood؛
  - الرقم 4 يعني أن الخوارزمية المتبعة هي خوارزمية Lesk.
- acceptedTags للدلالة على نوع الكلمات التي نرغب بمعالجتها ضمن النص سواء كانت فعل، اسم، صفة أو ظرف؛

accepted tags
NN=noun
VB=verb
JJ=adjective
RB=adverb

- withCoreference متحول منطقي يأخذ قيمة true عندما نريد إعادة الضمائر إلى أصلها أثناء معالجة النص المطلوب؛
- dbpediaPredicates لتحديد العلاقات التي نريد معالجتها والبحث فيها ضمن أنطولوجية dbpedia عندما يكون خيار searchInDbpedia مفعل؛

- leskWindowSize لتحديد حجم النافذة الأصغري الذي نريد أن نعالجه ضمن خوارزمية إزالة الغموض الرابعة lesk؛
- leskComparsionWay لتحديد طريقة المقارنة في خوارزمية lesk.

3 بنية ملف القواعد Rules File

- يحتوي هذا الملف على مجموعة القواعد الموضوعة من قبل خبراء بهدف حجب الصفحات بناء على محددات وغايات خاصة؛
- بعد وضع القواعد من قبل خبراء وباستخدام واجهة مساعدة يتم حفظ هذه القواعد في ملف القواعد الموجود عند ال Icap server والذي يقارن بين القواعد الموضوعة وبين نتائج معالجة النص المطلوب لمعرفة فيما إذا كانت الصفحة أم الملف سيحجب أم لا؛
- تتكون القاعدة في الحالة العامة من ثلاثة أقسام:
  - القسم الأول: مجموعة من ال domains مع نسبة وجود كل domain ضمن الصفحة؛
  - القسم الثاني: مجموعة من الكمات تفصل بينها علاقة (وَ &) ويتيح ال designer للخبير وضع كلمة مع معانيها بحيث نفصل بين كل كلمة ومعناها بفاصلة (,) وتعني سواء وردت هذه الكلمة أو أي كلمة من مرادفاتها؛
  - القسم الثالث الحدث Action الذي يجب تطبيقه عند تحقق القاعدة وقد يكون حجب الصفحة أو تمريرها B,U؛
- مثال:

Sexuality >= 0.2 & psychoanalysis >= 0.1 (sex, porn & girl) : u;

Sexuality >= 0.2 : b;

- القاعدة الأولى يقصد بها عند ورود المجال Sexuality بنسبة أكبر أو تساوي 0.2 والمجال psychoanalysis بنسبة أكبر أو تساوي 0.1 يكون قد تحقق القسم الأول للقاعدة، عند تحقق القسم الأول وورود الكلمة sex أو مرادفتها porn مع كلمة girl ستكون قد تحققت القاعدة كاملة عندها فقط يتم تنفيذ action القاعدة الأولى وهو عدم حجب الصفحة U؛
- القاعدة الثانية يقصد بها عند ورود المجال Sexuality بنسبة أكبر أو تساوي 0.2 تكون القاعدة قد تحققت وال action هو حجب الصفحة B؛
- ترتيب القواعد مهم جدا ففي حال كانت القاعدة غير محققة يتم الانتقال إلى القاعدة التي تليها حتما، وفي حال كانت القاعدة محققة والحدث هو عدم حجب الصفحة أو حجب الصفحة لا يتم الانتقال إلى القاعدة التي تليها؛
- العمليات التي من الممكن تطبيقها على المجالات هي <, >, =, <=, =>؛
- قمنا ببناء rule editor لبناء القواعد ومن ثم تخزينها في ملف القواعد؛
- هناك مفسر للقواعد rule parser يعمل على قراءة القواعد وتحليلها وفهمها.

مخدم البروكسي Smaz Proxy

3.1 Icap server

- هو زبون لخدمة الويب rmi client أي أن:

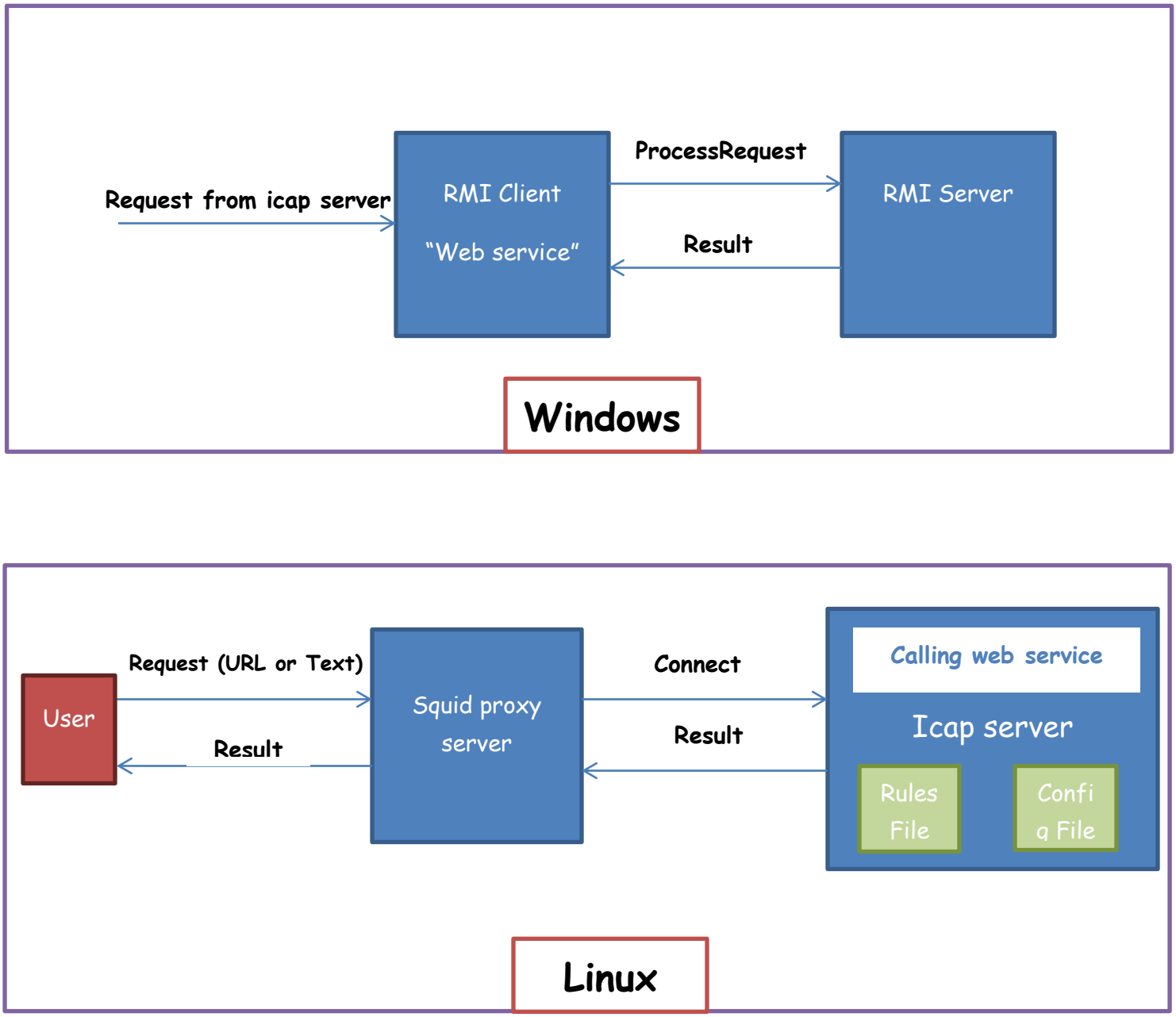
Icap server = client for web service (rmi client)

- عند طلب صفحة الويب من قبل الزبون يمر الطلب من خلال مخدم squid proxy server، ليتم تمرير الطلب بعدها إلى icap server الذي يحوي كود استدعاء خدمة الويب بعد الحصول على بارامترات الاستدعاء من خلال ملف الإعدادات الموجود عنده config file؛

- يقوم icap server بالاتصال بال rmi server ، ليقوم هذا الأخير بفهم النص دلاليا وتنفيذ تابع معالجة الطلب الذي يقدمه والحصول على كل domain مع نسبة وجوده ضمن النص، تُرد هذه النتيجة إلى icap server الذي يقوم بمقارنة نتيجة الطلب مع القواعد الموجودة عنده ضمن ملف القواعد rules file بالترتيب؛

- اعتمادا على هذه المقارنات يتم الحصول على نتيجة الطلب هل يتم حجبه أم لا، في حال كان هناك مطابقة مع قاعدة تقر بعد الحجب يتم تمرير الصفحة وعرضها للزبون ويتم التوقف عن مسح بقية القواعد، وفي حال كان هناك مطابقة مع قاعدة تقر بحجب الصفحة يتم حجب الصفحة وعدم عرضها على الزبون والتوقف عن مسح بقية القواعد.

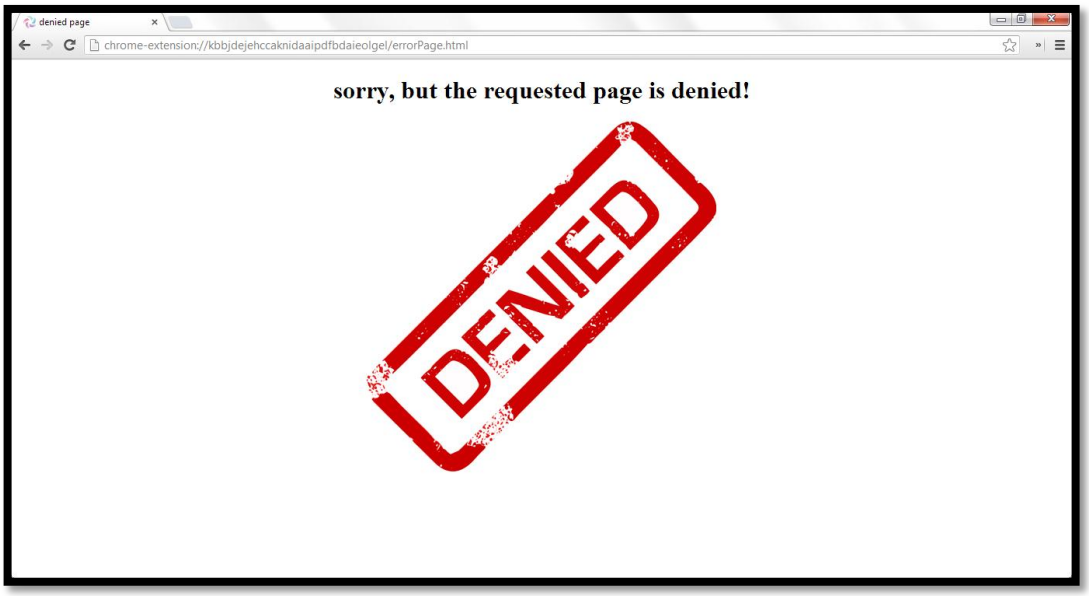
3.2 مخطط العمل وآلية الاتصال بين مكونات النظام



الشكل 61 مخطط العمل وآلية الاتصال بين المكونات

4 إضافة المتصفح Smaz Filter

- هنا يتم العمل ضمن المتصفح مباشرة من دون المرور بمخدم البروكسي icap وsquid ؛
- عند طلب الصفحة من خلال المتصفح أو تحديث الصفحة يتم معالجة الطلب من خلال خدمة الويب التي تعمل على فهم الصفحة ورد النتيجة، ليتم بعدها أخذ قرار حجب الصفحة أم لا من خلال القواعد المحددة مسبقا؛
- في حال كانت القرار هو حجب الصفحة يتم عرض الصفحة errorPage.html.



الشكل 62 صفحة الحجب

5 القارئ Smaz Reader

- تطبيق موبايل يعمل على تصنيف الأخبار إلى مجالات محددة وفقا للمحتوى؛
- يتم إضافة قناة أخبار من خلال تحديد الرابط الخاص بها؛
- يتم الحصول على الأخبار الموجودة ضمن الصفحة، وطلب خدمة الويب ومعالجة كل عناصر الأخبار؛
- نتيجة الاستعانة بهذه الخدمة يتم تصنيف الأخبار إلى مجالات معينة.

اختبار أداء Smaz Proxy

- لكي نتمكن من تحديد مدى فاعليه Smaz Proxy و قابليته للاستخدام الفعلي، قمنا بإجراء بعض الاختبارات التي تبين مدة تحميل بعض صفحات الانترنت مع استعمال Smaz Proxy وبدون استعماله؛
- قمنا بإجراء الاختبارات باستعمال حاسب (مخدم عتادي) يحمل المواصفات التالية:

  1- معالج من نوع Intel طراز Core I7 بتردد 2.2 GHz؛

  2- ذاكرة عشوائية بسعة 4 GB.

وباستعمال اتصال انترنت ADSL 1 mbps، و متصفح Mozilla Firefox 16، من دون تفعيل خيار البحث في DBpedia، ومن دون تفعيل خيار coreference، وباستخدام الخوارزمية WordNet tagged glosses لإزالة الغموض

فحصلنا على النتائج التالية :

زمن 2	زمن 1	الصفحة
29.30	13.64	www.huffingtonpost.com/news/syria
61.43	54.77	www.usnews.com/photos/violence-in-syria
23.54	13.65	www.insectidentification.org/cockroaches.asp
26.12	22.66	en.wikipedia.org/wiki/Politics_of_Israel
48.62	25.34	en.wikipedia.org/wiki/Sex
54.59	25.47	www.olympic.org/swimming-100m-backstroke-women

جدول 27 اختبار أداء SMAZ Proxy

حيث:

- زمن 1 : هو زمن تحميل الصفحة باستعمال مخدم proxy بدون اي اضافات؛
- زمن 2: هو زمن تحميل الصفحة باستعمال مخدم Smaz Proxy الذي يعتمد على الفلترة الدلالية؛
- وجميع الازمنة مقدرة بالثانية؛

نتيجة

بعد قيامنا بتجربة عدد لا بأس به من الصفحات و تحليل النتائج، لاحظنا ان استعمال Smaz Proxy يسبب زيادة حوالي 70٪ في مدة تحميل الصفحة و هذه نسبة مقبولة جدا نظرا للخدمات التي يقدمها.

ملاحظة

ضمن بيئة الاختبار الخاصة بنا تم تثبيت ال Proxy Server على نظام تشغيل Ubuntu Linux ضمن آلة افتراضية VirtualMachine

وهذا أدى الى بطئ في الأداء إلى حد ما، لكن عند يتم ثبيت ال Proxy على جهاز مستقل و بنظام تشغيل حقيقي سنلاحظ تحسن ملحوظ في السرعة.

6 البيئة البرمجية

6.1 لغة البرمجة المستخدمة

يعد اختيار لغة البرمجة المناسبة مرحلة مهمة، خاصة في الأنظمة المعقدة التي تتطلب شروطاً خاصة بها مثل وجود عدة أنظمة تشغيل أو عدة أنواع من نظم إدارة قواعد المعطيات، كل هذا وغيره يجعل اختيار لغة البرمجة المناسبة مرحلة مهمة من المستحق الوقوف عندها.

إن لغة الجافا هي لغة مناسبة جدا لهذا النظام لما تتمتع به من ميزات، أهمها الاستقلالية عن نظام التشغيل وعدة أسباب أخرى منها:

- لغة برمجة غرضية التوجه بشكل كامل؛
- المحمولية؛

- المقروئية وتنظيم اللغة؛
- القياسية؛
- الصيانية وسرعة التطوير؛
- أمن اللغة؛
- معالجة الاستثناءات والأخطاء؛
- دعم النياسب المتعددة؛
- دعم واجهة المستخدم البيانية؛
- الانتشار والتسويق والدعم الفني؛
- يوجد لها الكثير من المصادر على شبكة الإنترنت.

- Java

تم استخدام لغة البرمجة جافا في بناء التطبيقات الثلاثة الأساسية :

- Rmi server؛
- Rmi client؛
- Icap server.

وذلك باستخدام netbeans و eclipse enterprise edition.

- تم برمجة تطبيق الموبايل من خلال Java for android؛
- في تطبيق المتصفح تم استخدام JavaScript و HTML و CSS وملف توصيف لهذه الإضافة وهو مكتوب بلغة JSON؛
- في تطبيق الموبايل تم استخدام SQLite للتعامل مع قواعد المعطيات.

6.2 jsp web service

- تأمين خدمة الويب بشكل سهل.

6.3 ICAP

- Internet Content Adaption Protocol؛
- GreasySpoon؛
- يتيح هذا البروتوكول للمستخدم icap client بتمرير رسالة إلى مخدم icap server لإجراء عمليات المعالجة المطلوبة "adaptation"؛[11]
- يقوم المخدم بتنفيذ عمليات التحويل والمعالجة المطلوبة وإرسال النتيجة إلى الزبون، عادة ما تكون الرسالة معدلة؛
- غالبا ما تكون الرسالة هي طلب http request أو استجابة http response؛
- Icap عادة يستخدم لتحقيق مضاد فيروسات virus scanning وفلترة محتوى content filters.[12]

6.4 المكتبات المستخدمة

6.4.1 Apache Jena

- مكتبة مستخدمة في الجافا من أجل بناء تطبيقات الويب الدلالي؛
- Jena تقدم مجموعة من الأدوات ومكتبات الجافا التي تساعد في تطوير تطبيقات الويب الدلالي والمعطيات المترابطة Linked data؛
- Jena تتضمن:
  - API لقراءة ومعالجة وكتابة ثلاثيات ال RDF في ملفات XML ؛
  - القدرة على بناء أنطولوجيات باستخدام OWL والتعامل مع أنطولوجيات موجودة بصيغة RDF؛

- محرك بحث مزود بأخر توصيف للغة الاستعلام Sparql؛
- تخزين عدد كبير من ثلاثيات ال RDf ليتم استخدامها بشكل فعال لاحقا؛
- إمكانية نشر ثلاثيات ال RDF إلى تطبيقات أخرى باستخدام مجموعة من البروتوكولات بما فيها sparql؛
- محرك استدلال قائم على قواعد للتفكير مع مصادر البيانات RDF و OWL.

- تشبه مكتبة ال SemWeb و dotnetrdf الموجودة في ال C#.

6.4.2 Apache PDFBox

- أداة في الجافا مفتوحة المصدر تستخدم للتعامل مع مستندات ال PDF؛
- تتيح بناء مستندات PDF والتعامل مع مستندات موجودة مسبقا والقدرة على استخراج محتواها؛
- تم استخدام هذه المكتبة في المشروع لاستخراج محتوى مستندات ال PDF والحصول على النص الموجود فيها.

6.4.3 Apache POI

- أداة في الجافا مفتوحة المصدر تستخدم للتعامل مع مستندات ال office؛
- تتيح بناء مستندات office والتعامل مع مستندات موجودة مسبقا والقدرة على استخراج محتواها؛
- تم استخدام هذه المكتبة في المشروع لاستخراج محتوى مستندات ال office ولا سيما ملفات word والحصول على النص الموجود فيها.

6.4.4 Boilerpipe

- أداة في الجافا مفتوحة المصدر تستخدم للتعامل مع صفحات ال hltm؛
- تتيح بناء صفحات html والتعامل مع صفحات موجودة مسبقا والقدرة على استخراج محتواها؛

- تم استخدام هذه المكتبة في المشروع لاستخراج محتوى صفحات ال html والحصول على النص الموجود فيها.

6.4.5 Jsoup

- المكتبة تسمح بالتعامل مع صفحات HTML، حيث أنها تقوم بتحويل صفحة HTML إلى مجموعة من العقد والتي تمكن من البحث واستخراج المعلومات بسهولة، واستخراج العقد التي نريد بسرعة، كما تستطيع إنشاء صفحات أو إضافة عقد إليها أو تغيير عقد موجودة؛
- هذه المكتبة مشابهة HtmlAgilityPack لكنها تستخدم في الجافا، تم استخدام هذه المكتبة للحصول على النص المرافق للصور في صفحات ال html ، وذلك بهدف معالجة النص المحيط بالصور بغية حجب الصفحات التي تحوي صور يجب حجبها؛
- مثال عن استخدام هذه المكتبة حيث يتم الحصول على صفحة html ومن ثم الحصول على عقد محددة

```
Document doc = Jsoup.connect("http://en.wikipedia.org/").get();
Elements newsHeadlines = doc.select("#mp-itn b a");
```

6.4.6 JWI

- Java Wordnet Interface؛
- واجهة في الجافا تؤمن عمليات التعامل مع معجم ال wordnet؛
- تؤمن عمليات الوصول إلى المجموعات والتعامل معها بشكل مباشر والحصول على المجموعات synset ومعنى كل مجموعة gloss ومجموعة المترادفات وغير ذلك من العمليات؛

- تؤمن عمليات الحصول على كلمة معينة والحصول على الجذر المجرد الخاص بها lemma ومعناها بشكل سهل ومباشر؛
- مثال عن هذه العمليات:

```
public void runExample(){

    // construct the URL to the Wordnet dictionary directory
    String wnhome = System.getenv("WNHOME");

                                                    1

    String path = wnhome + File.separator + "dict";
    URL url = null;
    try{ url = new URL("file", null, path); }
    catch(MalformedURLException e){ e.printStackTrace(); }
    if(url == null) return;

                                                    2

   // construct the dictionary object and open it
   IDictionary dict = new Dictionary(url);
   dict.open();

                                                    3

   // look up first sense of the word "dog"
   IIndexWord idxWord = dict.getIndexWord("dog", POS.NOUN);
   IWordID wordID = idxWord.getWordIDs().get(0);
   IWord word = dict.getWord(wordID);
   System.out.println("Id = " + wordID);
   System.out.println("Lemma = " + word.getLemma());
   System.out.println("Gloss = " + word.getSynset().getGloss());

 }
```

1- الحصول على مكان وجود معجم wordnet ، المسار الموجود ضمنه المعجم؛

2- الحصول على غرض من المعجم ليتم التعامل معه؛

3- مجموعة من العمليات توضح آلية التعامل مع المعجم والذي تتيحه مكتبة jwi

a. الحصول على أول معنى لكلمة "dog"؛

b. الحصول على مميز هذه الكلمة؛

c. الحصول على الكلمة؛

d. الحصول على جذر الكلمة؛

e. الحصول على معنى المجموعة الموجود ضمنها كلمة "dog".

6.4.7 Stanford CoreNLP

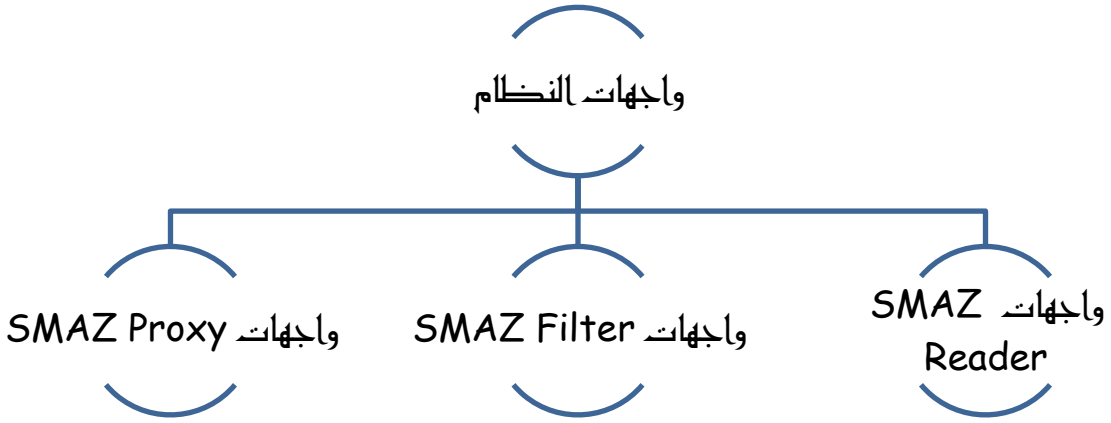
- مكتبة تؤمن لنا القدرة على إجراء عمليات معالجة اللغات الطبيعية:
  - تقسيم النص إلى مجموعة جمل وتدعى هذه العملية splitting؛
  - تقسيم الجملة إلى مجموعة كلمات وتدعى هذه العملية tokenizing؛
  - رد كل كلمة إلى أصلها الموجود ضمن معجم wordnet وتدعى هذه العملية lemmatizing؛
  - معرفة موقع كل كلمة ضمن النص هل هي فعل، فاعل، صفة أو ظرف وتدعى هذه العملية pos tagging؛
  - رد الضمائر إلى أصلها وتدعى هذه العملية coreferenceing.

6.5 SpotLight

- Spotlight هي أداة للتأشير التلقائي للنصوص على موارد ال DBpedia؛
- توفير حل لربط مصادر المعلومات غير المهيكلة لفتح سحابة البيانات المرتبطة من خلال DBpedia؛
- توفر web service دخلها النص وخرجها ال annotated text؛
- يمكن تمرير النص كامل وتحديد كلمات محددة للتأشير عليها؛
- قمنا بتمرير الجملة كاملة في محاولة لإزالة غموض الكلمات، حيث تقدم ال spotlight أيضا خدمة فك غموض النص الممرر لها.[10]

الفصل الثالث عشر

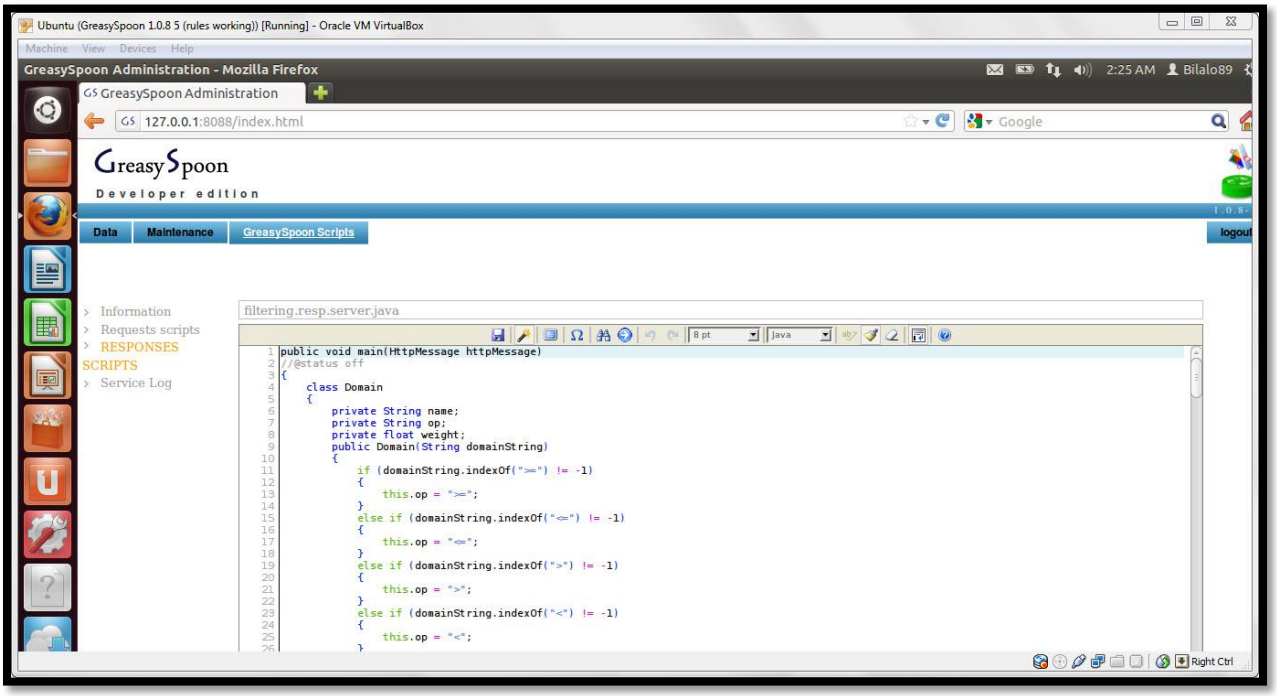
واجهات النظام



الشكل 63 النقاط الأساسية ضمن واجهات النظام

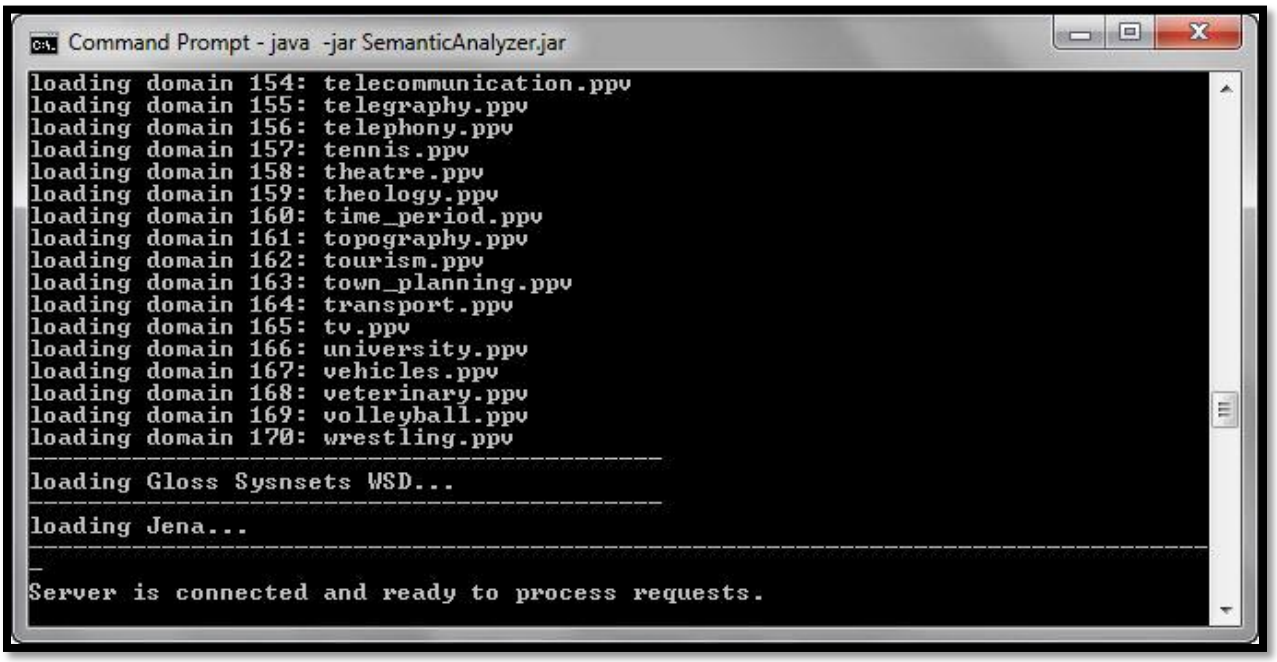
1 واجهات مخدم البروكسي Smaz Proxy

1.1 مخدم icap والبرمجة بداخله



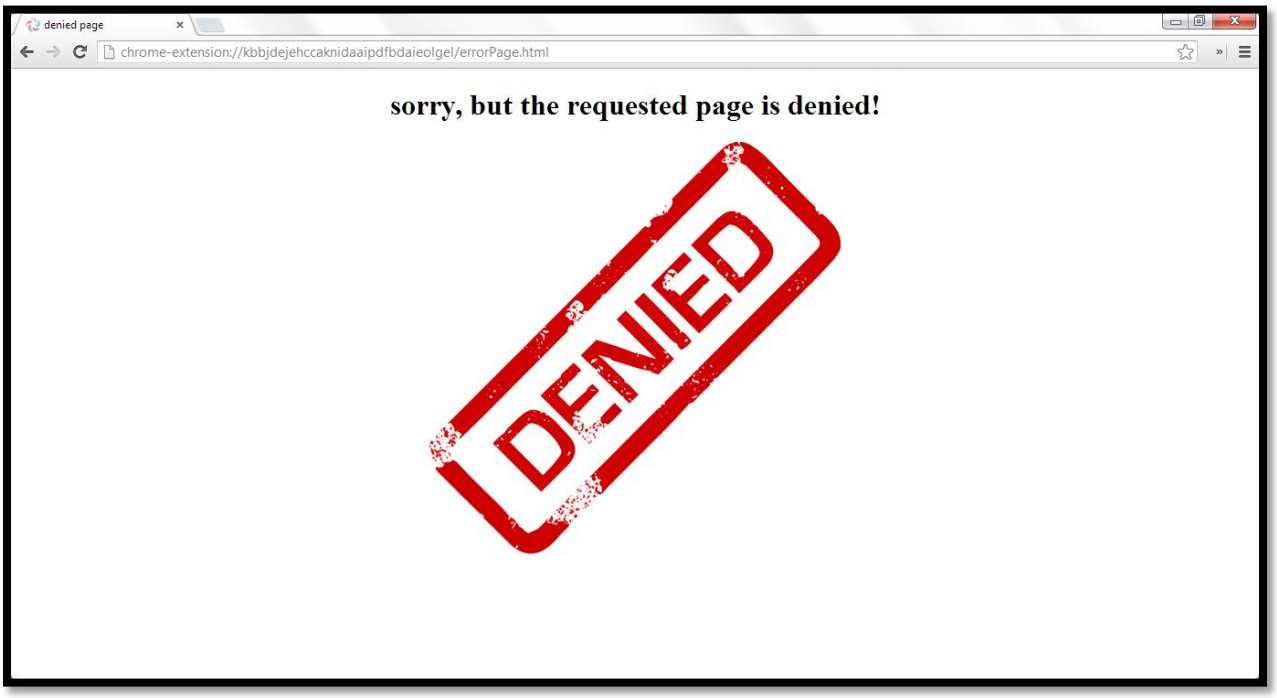
الشكل 64 مخدم icap والبرمجة بداخله

1.2 التأكد من أن المخدم متصل وقادر على معالجة الطلبات



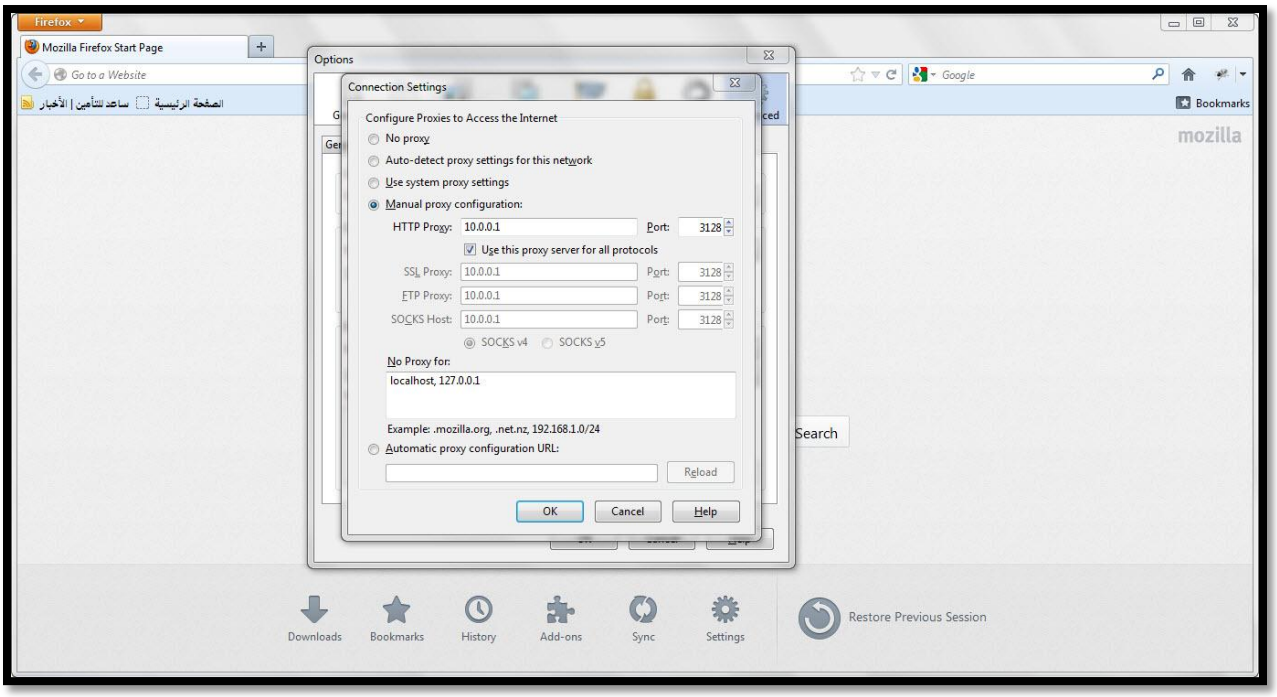
الشكل 65 التأكد من أن المخدم متصل وقادر على معالجة الطلبات

1.3 طلب صفحة ومعالجتها وحجبها



الشكل 66 طلب صفحة ومعالجتها وحجبها

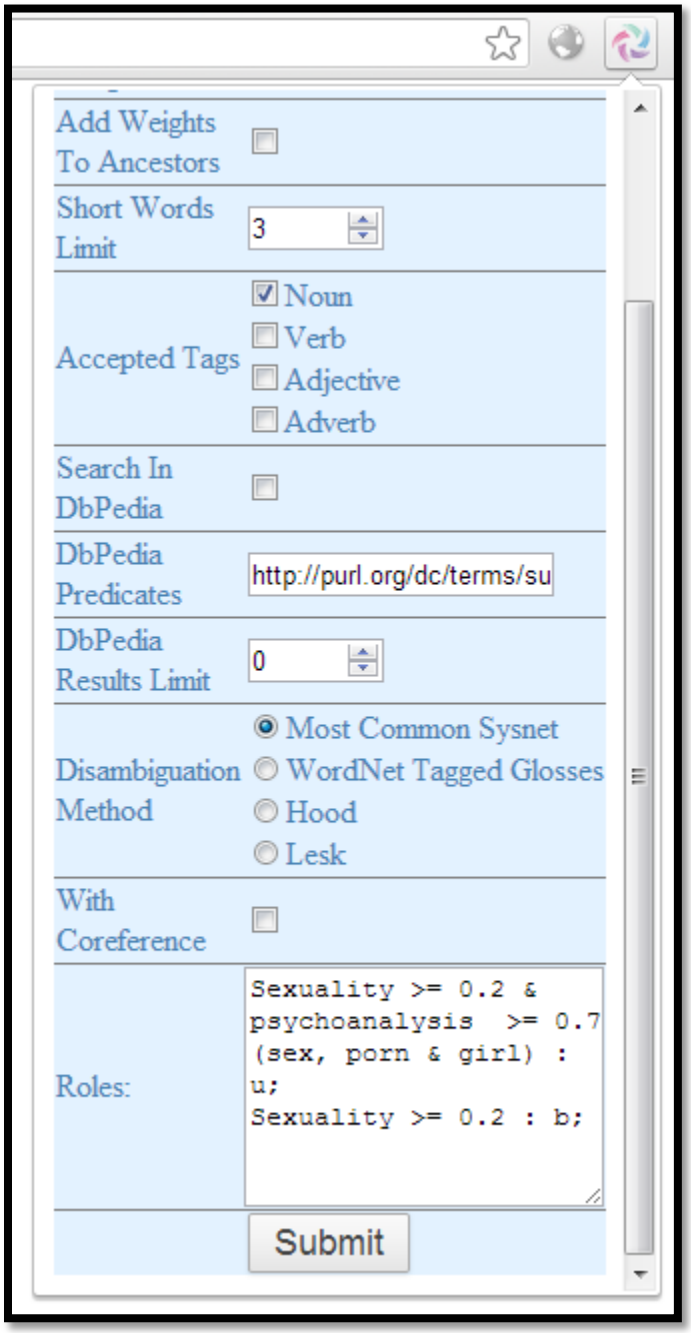
وضع إعدادات المخدم



الشكل 67 وضع إعدادات المخدم

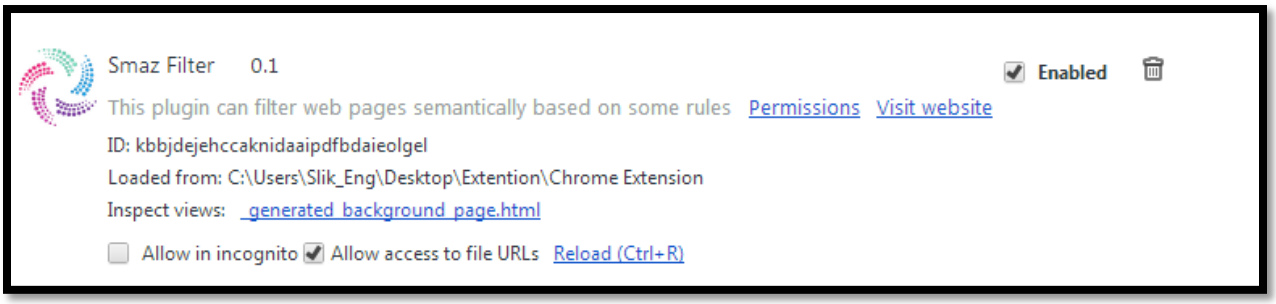
2 إضافة المتصفح Smaz Filtering

2.1 واجهة الإضافة التي يتم من خلالها وضع الإعدادات



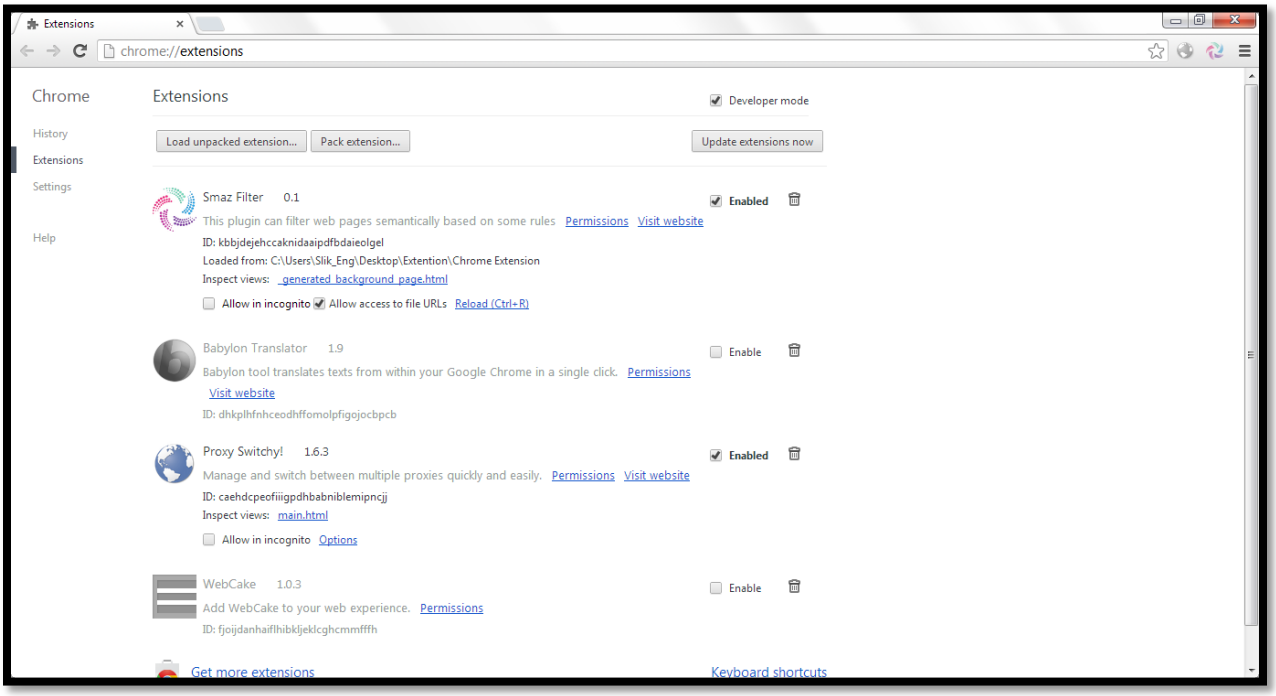
الشكل 68 واجهة الإضافة التي يتم من خلالها وضع الإعدادات

2.2 معلومات عن الإضافة



الشكل 69 معلومات عن الإضافة

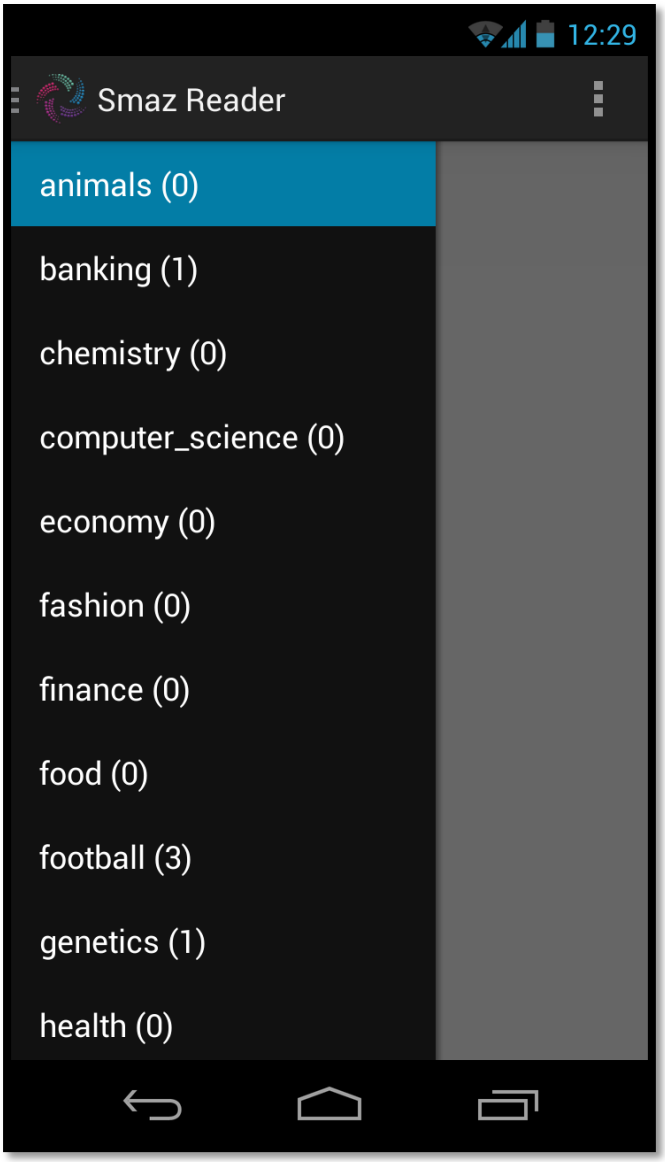
2.3 إضافة Smaz Filter إلى قائمة إضافات المتصفح



الشكل 70 إضافة Smaz Filter إلى قائمة إضافات المتصفح

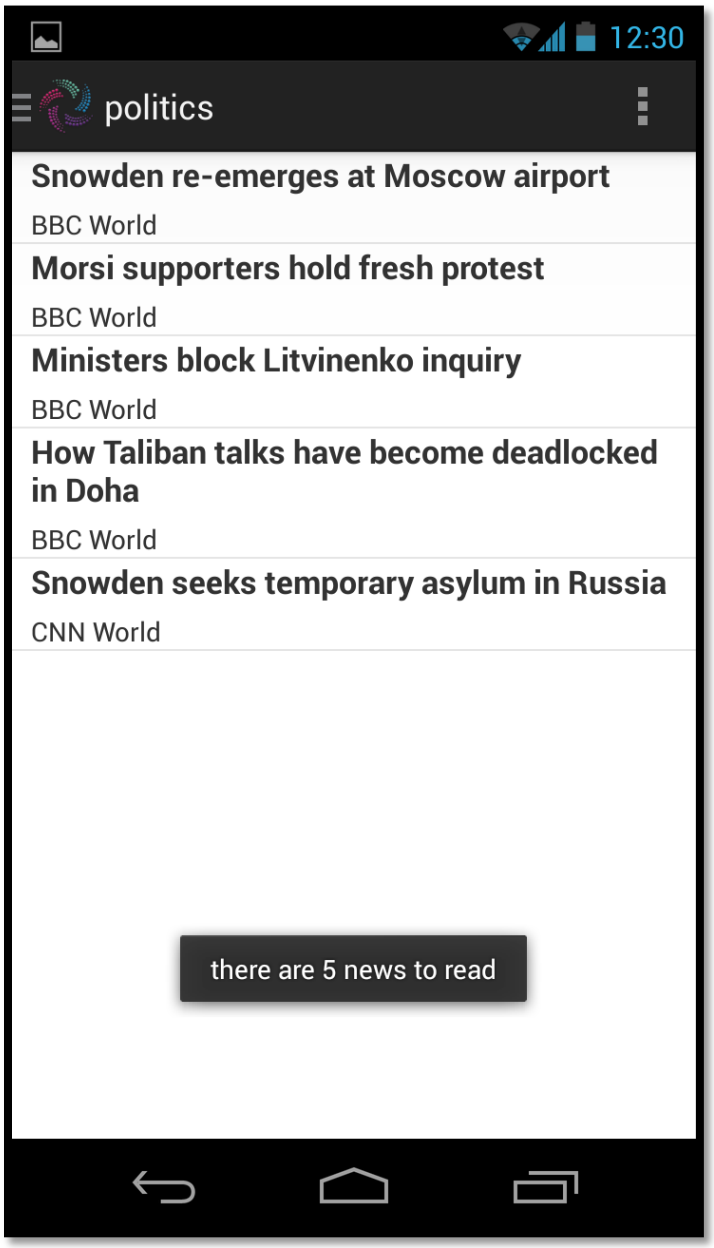
3 تطبيق الموبايل Smaz Reader

3.1 واجهة عرض المجالات المختارة مسبقا



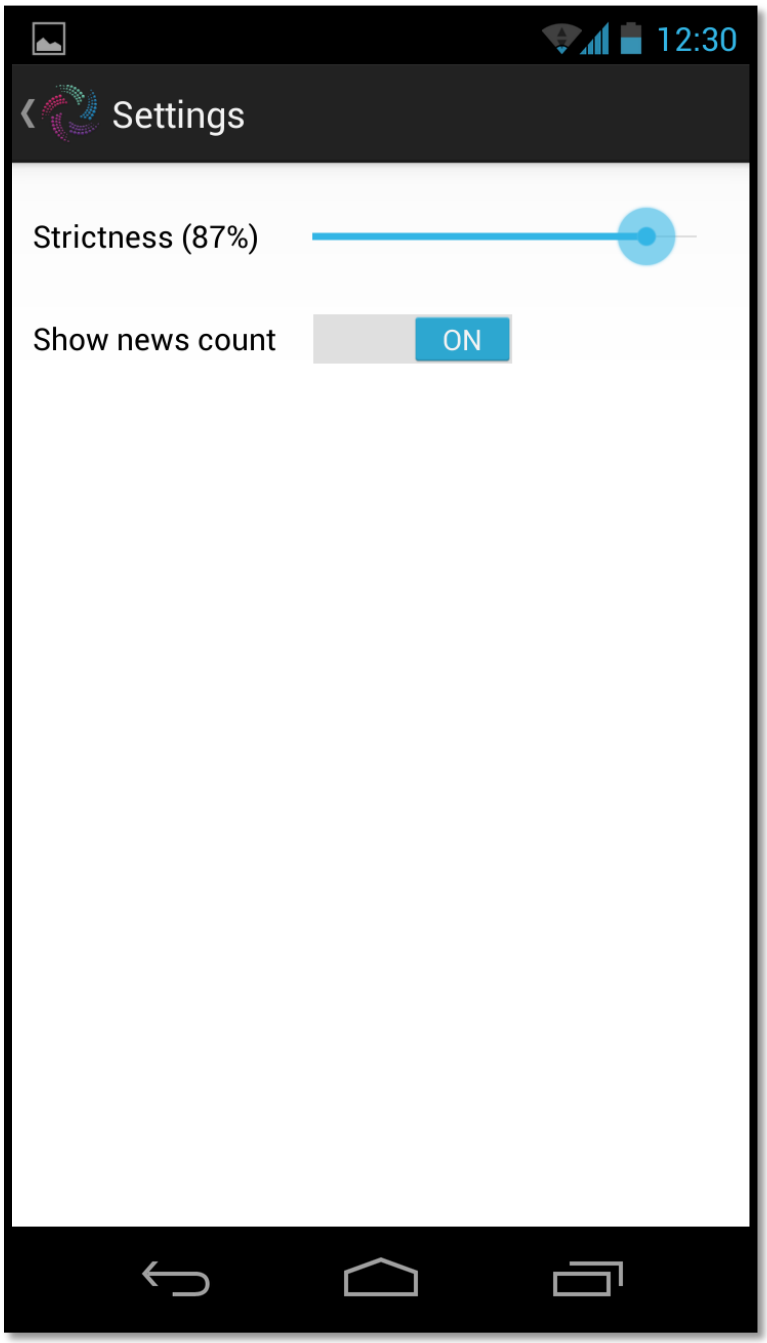
الشكل 71 واجهة عرض المجالات المختارة مسبقا

3.2 واجهة عرض الأخبار الخاصة بمجال معين



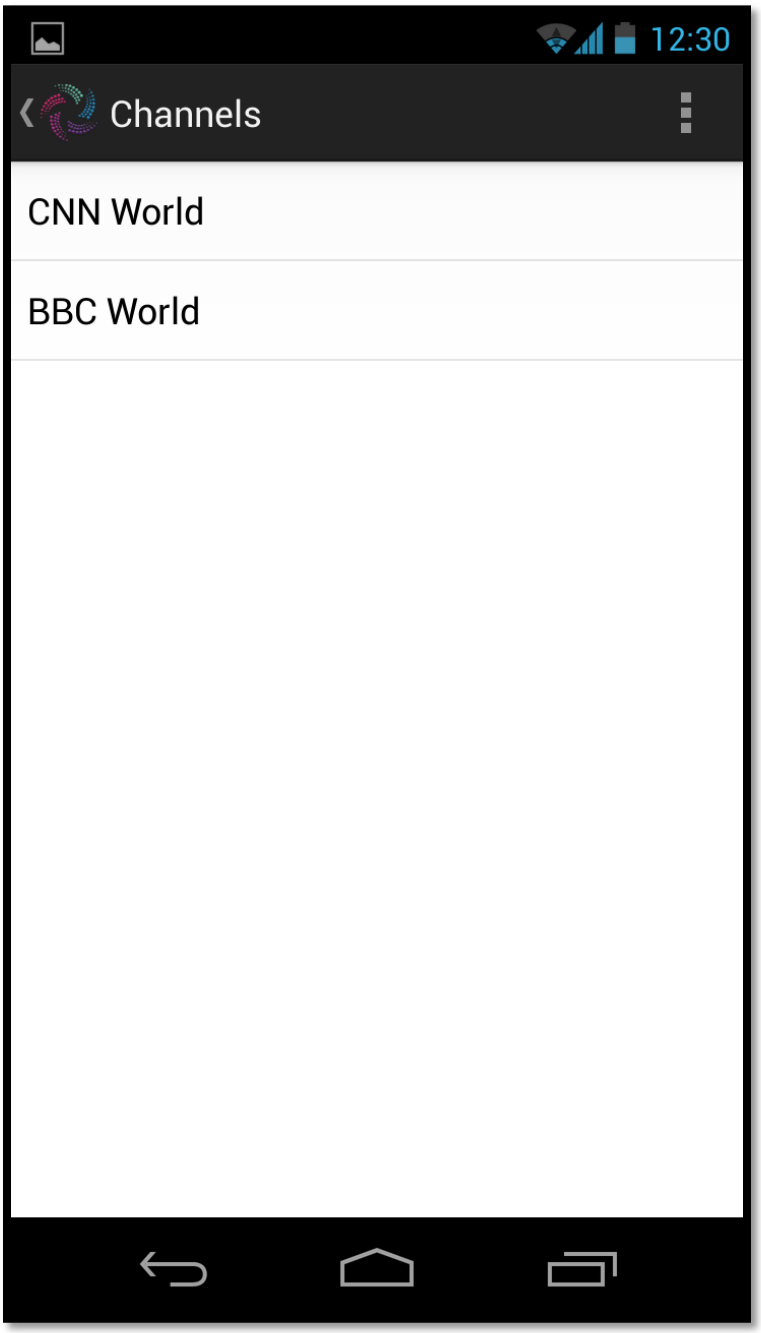
الشكل 72 واجهة عرض الأخبار الخاصة بمجال معين

3.3 واجهة إعدادات التطبيق



الشكل73 واجهة إعدادات التطبيق

3.4 واجهة عرض قنوات الأخبار

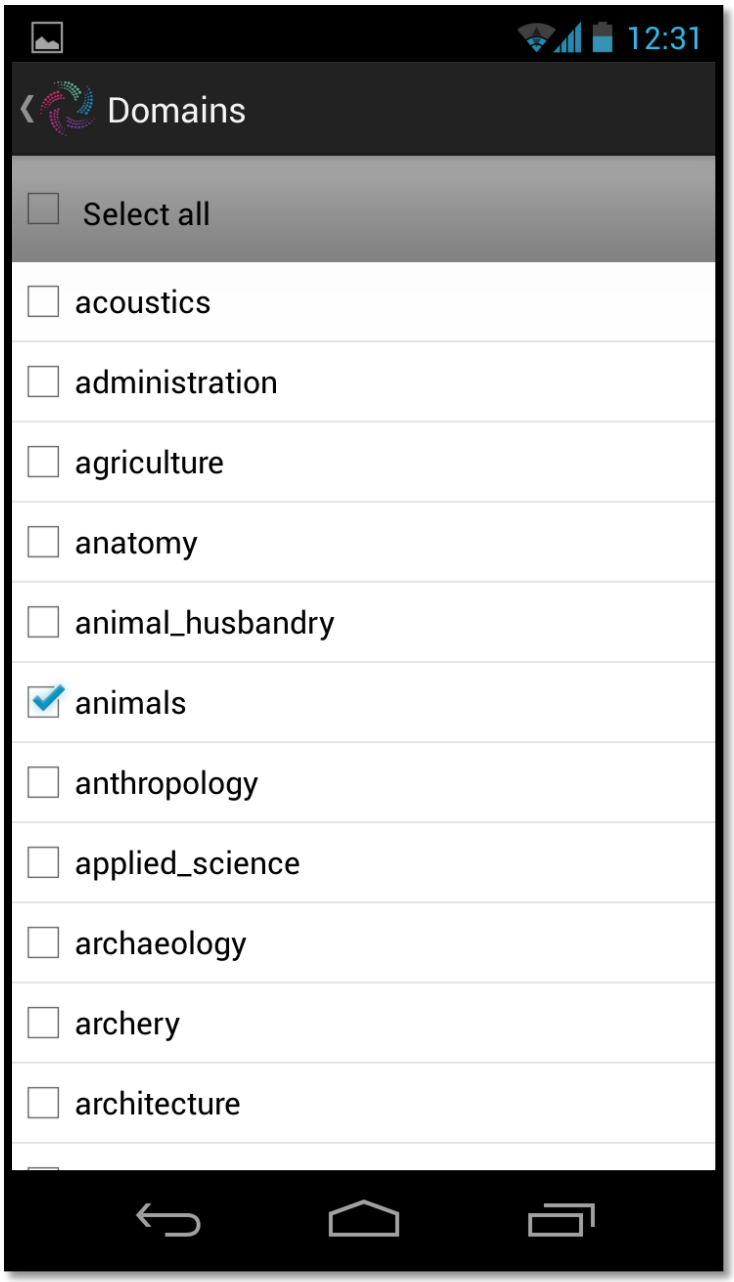


الشكل74 واجهة عرض قنوات الأخبار

3.5 واجهة تعديل قناة أخبار

الشكل 75 واجهة تعديل قناة أخبار جديدة

3.6 واجهة توضح اختيار المجالات المطلوب عرضها



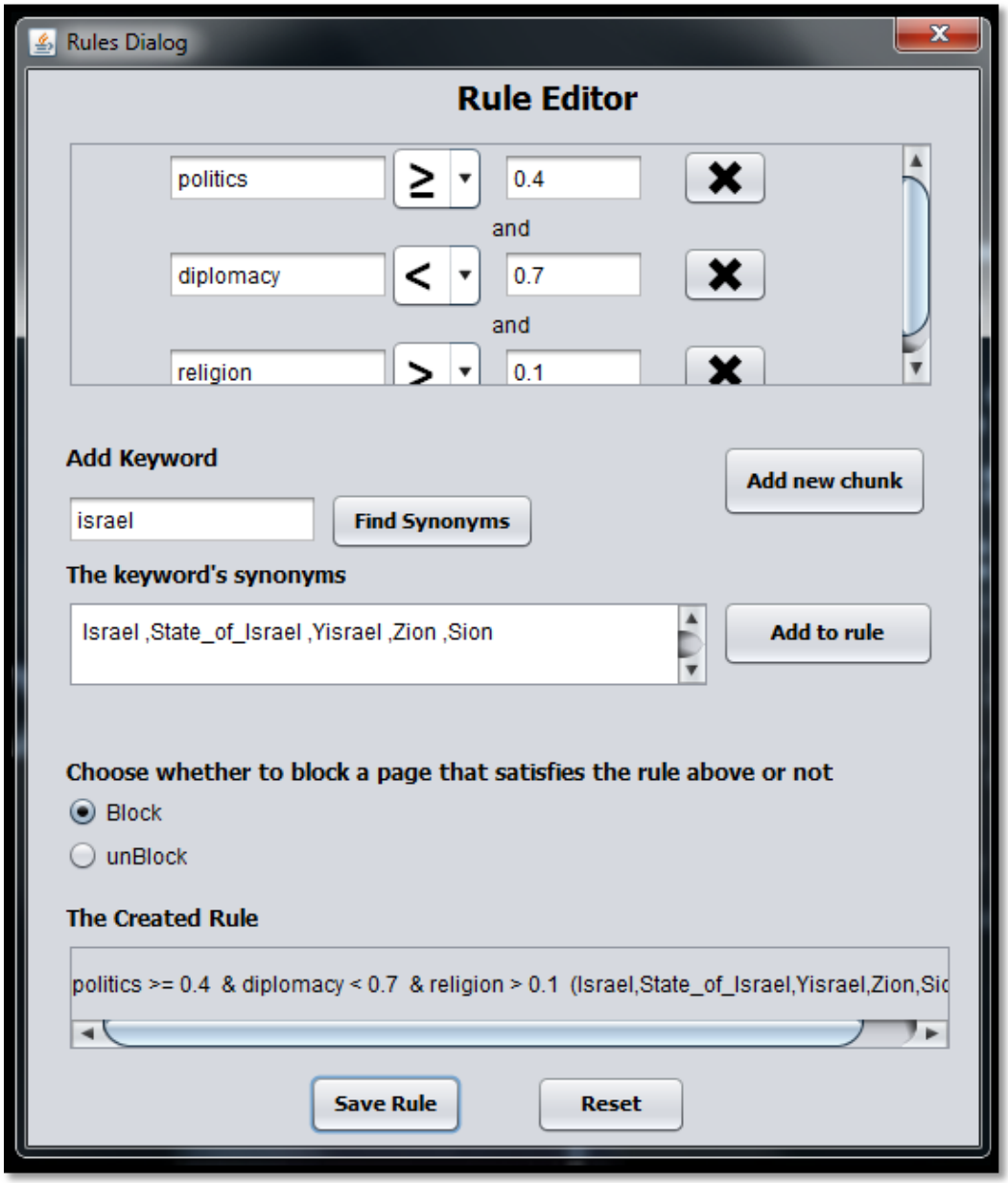
الشكل76 واجهة توضح اختيار المجالات المطلوب عرضها

3.7 واجهة توضع معلومات عن التطبيق

الشكل77 واجهة توضع معلومات عن التطبيق

4 محرر القواعد Rule editor

4.1 واجهة إضافة قاعدة

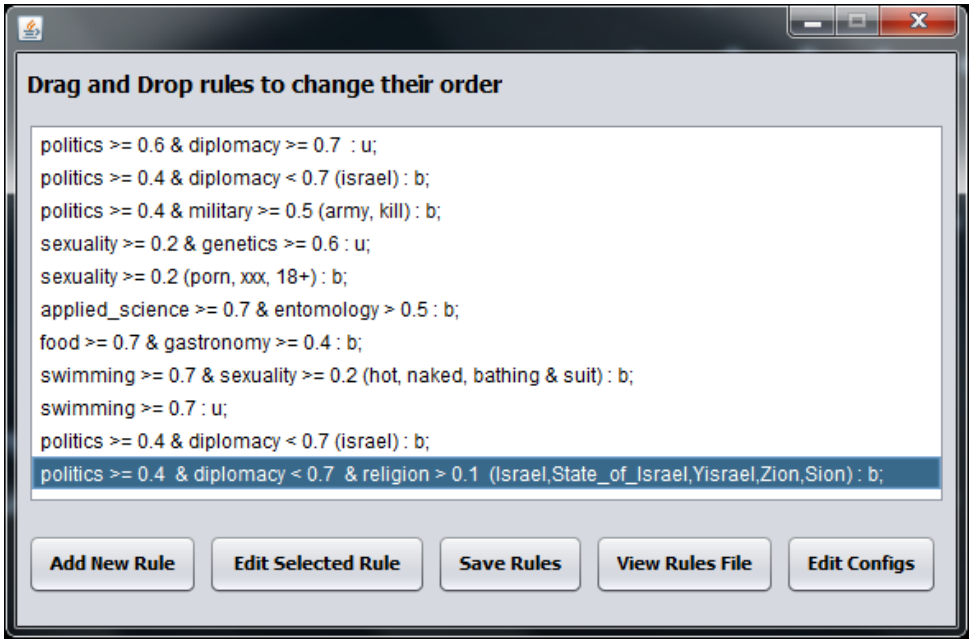


الشكل 78 واجهة إضافة قاعدة

4.2 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعة

الشكل 79 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعة

4.3 واجهة عرض القواعد الموجودة والتعديل عليها



الشكل80 واجهة عرض القواعد الموجودة والتعديل عليها

الباب السادس

مرحلة الاختبارات

1 مقدمة

يكمن التحدي الأكبر في التحقق من صحة عمل النظام بشكل صحيح وتأديته لجميع الوظائف المطلوبة، مما يقتضي إجراء كم كافي من الاختبارات، وبالطبع فإن الطريقة السليمة لاختبار أي نظام يجب أن تتم وفق منهجية صحيحة ومراحل متتالية للتأكد من صحة سير العمل، وبالتالي فقد تم تنفيذ مجموعة من الاختبارات عبر مراحل تطوير النظام ، وخلال مرحلة الاختبارات يجري التأكد من عدم ارتكاب اخطاء خلال مراحل تطوير النظام.

تتفاوت الأخطاء المكتشفة خلال هذه المرحلة في درجة خطورتها، لأن الخطأ الناتج عن مرحلة التحقيق البرمجي يعتبر من الاخطاء السهلة، بينما الخطأ الناتج عن مرحلة التحليل والتصميم يعتبر من أخطر الأخطاء وذلك لأنه يؤدي للعودة إلى تلك المراحل المبكرة لتصحيحه.

ولابد من التنويه بأن النظام يجب أن يحقق معايير الجودة الداخلية والخارجية.

1 الاختبارات عبر مراحل تطوير النظام

1.1 اختبار الواحدة البرمجية

هوالاختبار الذي قام به كل فرد من أفراد المشروع على الوحدة البرمجية التي طورها، ويكون كل فرد بذلك مسؤولا عن القيام بهذا النوع من الاختبارات.

1.1.1 اختبار التكامل

هوالاختبار الذي يتأكد من واجهات الوحدة البرمجية، وقدرتها على العمل مع بعضها، وقد تم هذا الاختبار عن الانتهاء من أجزاء متصلة من النظام، ولكن قبل أن يتم تجميع النظام ككل.

1.1.2 اختبار النظام

هوالاختبار الذي يتأكد من أن كل المتطلبات قد تحققت، وجرى هذا الاختبار عندما تم تجميع كل أجزاء النظام، ليتم التحقق من صحة عمل النظام ككل.

1.1.3 اختبارات الجودة

- **المعايير الخارجية**

1- الصحة

تعبر عن مدى ملائمة النظام لاحتياجات المستخدم، وقد تم تصميم النظام وفقا لما اتفق عليهليلبي الغرض الذي يم لأجله.

2- الوثوقية

تم مراعات الكثير من الأخطاء التي يمكن أن يرتكبها المستخدم من أجل الحصول على أعلى درجة من الوثوقية.

3- قابلية الصيانة

تم اعتماد مجموعة من الأمور في النظام لتأمين سهولة الصيانة وإمكانية اكتشاف الخطأ منها:

- لقد تم اعتماد نظام المجتزءات في تصميم النظام مما يسهل اكتشاف مكان الخطأ.
- القيام بعمليات الأرشفة لكل ما يدور في النظام من طلبات مما يسهل إمكانية متابعة العمل.

4- إعادة الاستخدام

بما أن النظام مكون من مركبات برمجية (مجتزءات)، فإن إعادة استخدام أحد المجتزءات ليس بالأمر الصعب.

- **<u>المعايير الداخلية</u>**

1- الاكتمال

تم تحقيق الوظائف المطلوبة من النظام بشكل كامل.

2- الاجتزائية

تم فصل أجزاء النظام المختلفة في مركبات مستقلة وظيفيا.

الباب السابع

الآفاق المستقبلية، الملحقات والمراجع

الفصل الرابع عشر

الآفاق المستقبلية

نأمل مستقبلا أن يتم العمل على تطوير المشروع والاستفادة أكثر من الخدمة التي نقدمها:

1- بناء محرك بحث وجعل نقطة الانطلاق هي المجالات domains، حيث تكون هذا المجالات بمثابة فهرس يتم من خلاله فهرسة صفحات الإنترنت وتوزيعها على المجالات التي تتبع إليها؛

2- دعم اللغة العربية من خلال تطبيق العمل على محتوى باللغة العربية، لم نقوم بهذه العملية لأن هذا العمل يتطلب أنطولوجية باللغة العربية وهذا الشي غير متوافر حالياً؛

3- الاستعانة بالمزيد من مصادر البيانات والربط مع أنطولوجيات أخرى وبالتالي زيادة الدقة في النتائج فالكلمات التي لا نجدها في wordnet كنا نبحث عنها في dbpedia، نطمح في المستقبل إلى الربط مع أنطولوجيات أكثر.

الفصل الخامس عشر

الملحق أ- مسرد المصطلحات

المصطلح باللغة الإنكليزية	اختصار المصطلح	المصطلح باللغة العربية
Children's Internet Protection Act	CIPA	قانون حماية الإنترنت للطفولة
meta tags		العلامات العامة
proxy server		مخدم بروكسي
chrome extention		إضافة على متصفح كروم
regular expressions		تعابير نظامية
filtering		فلترة
Phrase list		قائمة التعابير
Pornography		صور إباحية
Machine Learning		تعلم الآلة
Dynamic Real Time Rating	DRTR	التقييم الديناميكي في الوقت الحقيقي

add-on		إضافة
Domain		مجال
Grouping		تجميع
Meta Rules		قواعد عامة
Organization		تنظيم
Notification		إشعار
Anti-spam		مضاد لشيء غير مرغوب به
Wordnet		معجم
Wordnet Domains		مجالات المعجم
Synsets		مجموعات
Synonyms		مترادفات
Semantic		دلالي
Monosemous		كلمة لها معنى واحد
Polysemous		كلمة لها أكثر من معنى
Gloss		تفسير
Dbpedia		أنطولوجية
Resource Description Framework	RDF	إطار توصيف الموارد
metadata		معلومات عامة وصفية

Extensible Markup Language	XML	لغة التأشير الموسعة
Subject		فاعل
Predicate		فعل
Object		مفعول به
Dataset		بينات
Simple Protocol and RDF Query Language	SPARQL	لغة الاستعلام على ثلاثيات rdf
Normalization		تطبيع
Coreferencing		رد الضمير إلى أصله
Lemmatization		رد الكلمة إلى أصلها
Disambiguation		إزالة الغموض
Classification		تصنيف
corpus		فضاء
Stopwords		كلمات الغير حاملة للمعنى
Stemming		تجذير
Uniform Resource Identifier	URI	معرف الموارد المنتظم
Uniform Resource Locator	URL	محدد موقع الموارد المنتظم

web service		خدمة ويب
Combination		دمج
Channel		قناة
Remote Method Invocation	RMI	الاستدعاء البعيد للتوابع
Process		معالجة
Request		طلب
Client		زبون
Server		مخدم
Rule		قاعدة
rule editor		محرر القواعد
rule parser		مفسر القواعد
Web Ontology Language	OWL	لغة لبناء الأنطولوجيات
Java Wordnet Interface	JWI	واجهة للتعامل مع معجم wordnet
Splitting		تقسيم النص إلى جمل
Tokenizing		تقسيم الجمل إلى كلمات

جدول28 مسرد المصطلحات

الفصل السادس عشر

الملحق ب- فهرس المخططات

الشكل 1 مخطط غانت لمرحلة الدراسة النظرية 32

الشكل 2 مخطط غانت لمرحلة التحليل 32

الشكل 3 مخطط غانت لمرحلة التصميم 33

الشكل 4 مخطط غانت لمرحلة التحقيق 33

الشكل 5 مخطط غانت لمرحلة الاختبار 34

الشكل 6 النقاط الأساسية في الدراسة المرجعية 35

الشكل 7 النقاط الأسالسية في مشاريع الفلترة 37

الشكل 8 النقاط الأساسية في مشاريع الفلترة 38

الشكل 9 صورة توضح واجهة DansGuardian 39

الشكل 10 صورة توضح واجهة K9 42

الشكل 11 صورة توضح واجهة OpenDNS 45

الشكل 12 صورة توضح التعامل مع openDNS 46

الشكل 13 صورة توضح واجهة We-Blocker 47

الشكل 14 صورة توضح واجهة anti-porn 48

الشكل 15 صورة توضح واجهة squid 50

الشكل 16 صورة توضح واجهة safe-squid 53

الشكل 17 آلية العمل DansGuardian + squid مرحلة 1 57

الشكل 18 آلية العمل DansGuardian + squid مرحلة 2 57

الشكل 19 آلية العمل DansGuardian + squid مرحلة 3 58
الشكل 20 آلية العمل DansGuardian + squid مرحلة 4 58
الشكل 21 النقاط الاساسية في فلترة البريد الإلكتروني 59
الشكل 22 واجهة outlook مرحلة 61
الشكل 23 واجهة outlook مرحلة 2 62
الشكل 24 واجهة outlook مرحلة 3 64
الشكل 25 واجهة outlook مرحلة 4 65
الشكل 26 واجهة outlook مرحلة 5 66
الشكل 27 واجهة outlook مرحلة 6 67
الشكل 28 واجهة outlook مرحلة 7 68
الشكل 29 واجهة outlook مرحلة 8 69
الشكل 30 النقاط الأساسية في مرحلة الدراسة النظرية 73
الشكل 31 مثال توضيحي عن العلاقات الدلالية في wordnet 76
الشكل 32 مثال توضيحي عن الارتباطات في wordnet 78
الشكل 33 مثال عن ثلاثية rdf 82
الشكل 34 مثال توضيحي عن ثلاثيات rdf 82
الشكل 35 النقاط الأساسية في تحليل وفهم المحتوى 85
الشكل 36 مثال عن غموض كلمة 88
الشكل 37 خوارزميات إزالة الغموض 89
الشكل 38 الدمج بين المجموعات 93
الشكل 39 أسلوب المقارنة بين المجموعات 94
الشكل 40 حساب رصيد كل تركيبة 98
الشكل 41 مثال عن إيجاد جذر hood 100
الشكل 42 حالات تحديد الجذر في خوارزمية hood 102
الشكل43 النقاط الأساسية ضمن مرحلة التحليل 111

الشكل 44 مخطط حالات الاستخدام لمخدم البروكسي 118
الشكل 45 مخطط حالات الاستخدام لقارئ الأخبار..... 119
الشكل 46 إضافة قاعدة إلى قواعد الفلترة..... 136
الشكل 47 حذف قاعدة..... 137
الشكل 48 عرض الأخبار التابعة لمجال معين 137
الشكل 49 إضافة قناة أخبار..... 138
الشكل 50 وضع إعدادات للتطبيق 138
الشكل51 مخطط الصفوف المفاهيمي لمخدم البروكسي 140
الشكل52 مخطط الصفوف المفاهيمي لقارئ الأخبار..... 145
الشكل53 النقاط الأساسية ضمن مرحلة التصميم 147
الشكل 54 مخطط صفوف مخدم البروكسي..... 149
الشكل 55 مخطط صفوف قارئ الأخبار 156
الشكل56 مخطط قاعدة المعطيات..... 166
الشكل 57 النقاط الأساسية ضمن مرحلة التحقيق..... 167
الشكل 58 النقاط الأساسية في تحقيق النظام 170
الشكل59 مخطط يوضح آلية عمل خدمة الويب 171
الشكل 60 آلية عمل RMI..... 172
الشكل 61 مخطط العمل وآلية الاتصال بين المكونات..... 180
الشكل 62 صفحة الحجب 181
الشكل 63 النقاط الأساسية ضمن واجهات النظام..... 193
الشكل 64 مخدم icap والبرمجة بداخله 194
الشكل 65 التأكد من أن المخدم متصل وقادر على معالجة الطلبات 195
الشكل 66 طلب صفحة ومعالجتها وحجبها 196
الشكل 67 وضع إعدادات المخدم..... 197
الشكل 68 واجهة الإضافة التي يتم من خلالها وضع الإعدادات..... 198

الشكل 69 معلومات عن الإضافة 199
الشكل 70 إضافة Smaz Filter إلى قائمة إضافات المتصفح 199
الشكل 71 واجهة عرض المجالات المختارة مسبقا 200
الشكل 72 واجهة عرض الأخبار الخاصة بمجال معين 201
الشكل73 واجهة إعدادات التطبيق 202
الشكل74 واجهة عرض قنوات الأخبار..... 203
الشكل 75 واجهة تعديل قناة أخبار جديدة 204
الشكل76 واجهة توضح اختيار المجالات المطلوب عرضها 205
الشكل77 واجهة توضع معلومات عن التطبيق 206
الشكل 78 واجهة إضافة قاعدة 207
الشكل 79 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعة 208
الشكل80 واجهة عرض القواعد الموجودة والتعديل عليها..... 209

الفصل السابع عشر

الملحق ج- فهرس الجداول

جدول2 إحصائيات عن معجم wordnet 79

جدول 3 إحصائيات عن أنطولوجية DBpedia 83

جدول4 مقارنات بين الخوارزميات الاربعة في إزالة الغموض..... 106

جدول 5 وضع إعدادات النظام..... 120

جدول 6 تعديل إعدادات النظام 121

جدول 7 إضافة قاعدة فلترة 122

جدول8 تعديل قاعدة فلترة..... 124

جدول9 طلب صفحة ويب..... 125

جدول 10 حذف قاعدة فلترة..... 126

جدول 11 عرض محتوى ملف القواعد..... 127

جدول 12 عرض محتوى ملف الإعدادات..... 128

جدول13 عرض الأخبار التابعة لمجال معين..... 129

جدول14 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها 130

جدول 15 وضع إعدادات التطبيق..... 131

جدول 16 عرض المجالات المتواجدة 132

جدول 17 إضافة قناة أخبار..... 133

جدول 18 عرض القنوات المتواجدة 134
جدول 19 حذف قناة 134
جدول20 وضع إعدادات النظام 135
جدول 21 توصيف الصفوف المفاهيمية لمخدم البروكسي 144
جدول22 توصيف الصفوف المفاهيمية للقارئ 146
جدول 23 توصيف الكيان rssChannel 163
جدول 24 توصيف الكيان Domain 164
جدول 25 توصيف الكيان RssItem 164
جدول 26 توصيف الكيان DomainRssItem 165
جدول 27 اختبار أداء SMAZ Proxy 183
جدول28 مسرد المصطلحات 224

الفصل الثامن عشر

الملحق د - المراجع

[1] Informations about dansguardian. [Online] *http://dansguardian.org*.

[2] Informations about K9.[Online] *http://en.wikipedia.org/wiki/K9_Web_Protection*.

[3] Informations about opendns. [Online] *www.opendns.com*.

[4] Informations about Squid. [Online] *en.wikipedia.org/wiki/Squid*.

[5] Informations about quintolabs. [Online] *http://issues.quintolabs.com/trac/quintolabs_qlicap/wiki/BlockSites.*

[6] Informations about quintolabs. [Online] *http://www.quintolabs.com.*

[7] Informations about safesquid. [Online] *www.safesquid.com.*

[8] Informations about wordnet. [Online] *http://wordnet.princeton.edu.*

[9] Informations about wordnet domains. [Online] *http://wndomains.fbk.eu.*

[10] Informations about spotlight - Linking to dbpedia. [Online] *http://spotlight.dbpedia.org.*

[11] Informations about Icap. [Online] *http://en.wikipedia.org/wiki/Internet_Content_Adaptation_Protocol.*

[12] Informations about features of Icap. [Online] *http://wiki.squid-cache.org/Features/ICAP.*

[13] Informations about features of word sense disambiguation. [Online] *http://julian.eti.pg.gda.pl/publikacje/wsd.pdf.*

[14] Informations about features of word sense disambiguation - gloss taged algorithm. [Online] *http://wordnet.princeton.edu/glosstag.html.*

[15] Informations about features of word sense disambiguation - gloss taged algorithm. [Online] *http://kask.eti.pg.gda.pl/semagloss/index.html.*

[16] Informations about wordnet. [Online] *https://ptl.sys.virginia.edu/ptl/members/matthew gerber/software#wordnet.*

[17] Satanjeev Banerjee. *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*, 2002.

[18] Jonas EKEDAHL, **koraljka Golub**. *Word sense disambiguation using WordNet* and the Lesk algorithm.

[19] Kostas Fragos, **Yannis Maistros**, **Christos Skourlas**. *Word Sense Disambiguation Using WordNet Relations*.