الفصل الحادي عشر

استكشاف المعلومات الموزعة أو اللامركزية

تعد عملية استكشاف معلومات مبثوثة وموزعة عبر العديد من النظم الآلية إحدى قوى التحدي التي تواجه عملية التشغيل المتداخل، فالمكتبات الرقمية على مستوى العالم تديرها جهات متعددة، وبأساليب إدارية متنوعة، وبتوجهات مختلفة، بناءً على طبيعة مجموعاتها، وعلى التقنيات التي يتم تطبيقها. كما أن قلة من هذه المكتبات هي التي تمتلك نسبة صغيرة من المواد التي يطلبها المستفيدون بكثرة؛ ومن ثم فإن المستفيدين يضطرون إلى طرق أبواب مختلف المجموعات والإفادة من الخدمات التي تقدمها كثير من المصادر المختلفة، وعليه كيف يستطيع المستفيد استكشاف المعلومات المبثوثة عبر هذه المصادر الكثيرة المحتملة، وكيف يمكنه الوصول إليها ؟

إن مصطلح التطبيقات الحاسوبية الموزعة أو اللامركزية Distributed Computing، مصطلح عام يعبر عن جميع الجوانب الفنية المرتبطة بعملية التنسيق بين الحاسبات الآلية التي تعمل مستقلة بعضها عن بعض حتى تتمكن من تقديم خدمة متناسقة. وتتطلب التطبيقات الحاسوبية الموزعة أن يكون هناك قدر مشترك من المعيارية الفنية أو التوحيد بين الحاسبات المختلفة، فقد يرغب المستفيد عند إجراء عملية البحث الموزعة - على سبيل المثال - أن يبحث في عدة مجموعات مستقلة بعضها عن بعض من خلال استفسار واحد، ثم يقوم بعد ذلك بمقارنة النتائج، ثم اختيار أكثر هذه النتائج قرباً من موضوع الاستفسار، ثم استرجاع المواد التي يقع عليها الاختيار من

بين هذه المجموعات. وعلاوة على ضرورة توافر صفة المعيارية الخاصة بعملية المشابكة، فإن هذه العملية - أي البحث الموزع - تتطلب توافر طريقة ما للتعرف إلى المقتنيات، ومعرفة ضوابط صياغة الاستفسارات وأساليب تقديمها، ومعرفة أساليب استرجاع النتائج، وطرق الحصول على المواد التي تم استكشافها، وقد تكون هذه المعايير رسمية معتمدة من قبل الجهات الرسمية المتخصصة في إصدار المواصفات القياسية، أو قد تكون معايير محلية طورتها مجموعة صغيرة من الجهات المتعاونة، أو مجرد اتفاقيات تنظم تداول منتجات تجارية معينة.

ولعل التوجه المناسب في هذا الصدد يتمثل في تطوير مجموعة شاملة من المعايير التي يمكن أن تتبناها جميع المكتبات الرقمية، ومع ذلك، فإن هذه الفكرة تخفق في معرفة تكاليف تبني هذه المعايير الشاملة والالتزام بها لاسيما في ظل هذه التغيرات المتسارعة التي نشهدها هذه الأيام؛ فالمكتبات الرقمية في تغير متواصل، وكل مكتبة تسعى لتطوير مجموعاتها وخدماتها ونظمها بحيث لا تتشابه مكتبتان في هذا التوجه، كما ينظر إلى أن عملية استبدال جزء من النظام أو تغييره لدعم معيار جديد هي عملية مضيعة للوقت، حيث يمكن أن تظهر نسخة جديدة من المعيار قبل أن تكتمل عملية الاستبدال هذه، أو ربما يكون الناس قد انصرفوا عن تطبيقه، واتجهوا نحو طريق آخر؛ فعملية التوحيد القياسي الكامل ما هي إلا سراب.

أما بالنسبة لعملية التشغيل المتداخل، فمن الواضح أن المكتبات الرقمية تواجه تحدياً كبيراً يتمثل في وجود نظم حاسوبية موزعة تعمل في سياق عالم تسوده حاسبات تعمل منفصلة بعضها عن بعض، وغير متوافقة من

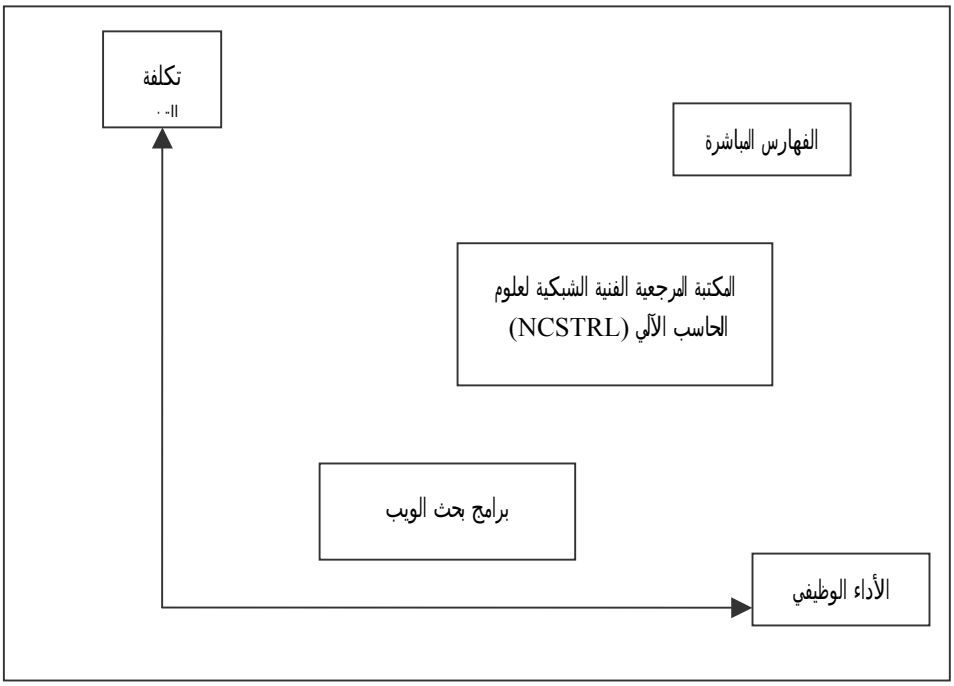
الناحية الفنية. ولكي يتسنى تبادل الرسائل فيما بين هذه الحاسبات، يتطلب الأمر وجود صيغ وبروتوكولات ونظم أمن تؤمن ذلك، كما أن الأمر يحتاج كذلك إلى وجود اتفاقيات تختص بدلالة الألفاظ للمساعدة في ترجمة تلك الرسائل وتفسيرها. ولكن يظل التحدي الأساسي متمثلاً في إيجاد طرق تحفز المكتبات الرقمية المنفصلة عن بعضها عن بعض على التعاون وعلى تضافر الجهود.

لا شك أن تبني طرق مشتركة يمنح المكتبات الرقمية مزيداً من القدرة على أداء وظائفها، غير أن لهذا النمط من التبني تكاليفه التي قد يكون بعضها مالياً كتكاليف شراء التجهيزات والبرمجيات، وتعيين الموظفين وتدريبهم؛ وقد تكون معظم هذه التكاليف في أحيان كثيرة مرتبطة بالنواحي التنظيمية، ونادراً ما يمكن تغيير أحد جوانب المكتبات الرقمية بمعزل عن الجوانب الأخرى. وإن طرح معيار جديد يستلزم إجراء تغييرات متداخلة في النظم الحالية، كما يتطلب كذلك تغييرات أخرى في سير العمل وتدفقه، وفي طبيعة العلاقات مع الموردين، وربما يتطلب تغييرات أخرى أكثر من ذلك.

ويصور الشكل رقم (١١-١) أحد النماذج التصورية التي يمكن الاستفادة منها عند التفكير في تدشين عملية التشغيل المتداخل، ويتم استخدام النموذج في هذه الحالة للمقارنة بين ثلاث طرق مختلفة من طرق البحث الموزع أو اللامركزي؛ حيث يشير المحور الأفقي في الشكل إلى الأداء الوظيفي الذي تقوم به الطرق المختلفة، في حين يشير المحور الرأسي إلى تكلفة تبني هذه الطرق. ويلاحظ أن الطريقة المثالية هي التي تتجسد في الأسفل يمين الشكل؛ حيث يتم تحقيق أعلى قدر من الأداء الوظيفي وبأقل تكلفة، فبرامج بحث الويب

تتمتع بقدرة متوسطة على أداء الوظائف، لكنها شائعة الاستخدام بسبب انخفاض تكاليف تبنيها، أما الفهارس المباشرة المعتمدة على معايير الفهرسة المقروءة آلياً المعروفة بمارك MARC، وعلى المعيار Z39.50، فإنها تتمتع بقدرة كبيرة على أداء الوظائف، لكنها محدودة الاستخدام بسبب معاييرها المعقدة، أما نظام المكتبة المرجعية الفنية الشبكية لعلوم الحاسب Networked Computer Science Technical Reference Library (NCSTRL) (المذكور في اللوحة ١١-٤) فيقع في مرحلة وسط بين هاتين الطريقتين.

الشكل رقم (١١-١)
مقارنة بين ثلاث من استراتيجيات البحث الموزع أو اللامركزي



وتقع الطرق التي يمكن استخدامها لأغراض عملية التشغيل المتداخل في

ثلاث فئات واسعة على النحو التالي:

- تتمتع معظم الطرق التي يشيع استخدامها على نطاق واسع لأغراض عملية التشغيل المتداخل اليوم بإمكانيات متوسطة على الأداء الوظيفي، وتكلفة تطبيق منخفضة. وتعد المعايير الأساسية المرتبطة بالويب المتمثلة في لغة ترميز النصوص الفائقة (HTML)، وبروتوكول نقل الملفات (HTP)، والمحدد الموحد لموقع المصدر (URL)، من النماذج التي تتمتع بهاتين الخاصتين. ومن الملاحظ أن البساطة التي اتسمت بها هذه المعايير قد أدت إلى تبنيها على نطاق واسع، غير أنها تحد من كفاءتها الوظيفية.

- توفر بعض الخدمات التي تستخدم أحدث التقنيات قدرة كبيرة على الأداء الوظيفي لكن تكلفة تبنيها عالية جداً، وخير مثال على ذلك معيار Z39.50، واللغة المعيارية الموحدة للترميز (SGML). وعادة ما يشيع استخدام هذه الأساليب في مجتمعات محددة، حيث تزداد أهمية الأداء الوظيفي، لكن تكلفة تبني هذه الطرق تعيق من تطبيقها في مجتمعات أكبر.

- تعد كثير من التطورات الحالية في مجال المكتبات الرقمية محاولات إيجاد أرضية مشتركة تتمثل في: زيادة القدرة على أداء الوظائف بتكلفة تطبيق متوسطة. وتشمل الأمثلة على ذلك كل من: معيار دبلن كور (Dublin Core)، ولغة الترميز الموسعة (XML)، ونظام اليونيكود (Unicode)؛ حيث حرص المصممون في جميع النماذج السابقة على توفير طرق معقولة التكلفة عند تبنيها؛ فدبلن كور مثلاً يسمح دائماً بأن تكون جميع الحقول اختيارية، كما أن نظام اليونيكود قدم صيغة يو تي إف-٨ التي تسمح بقبول البيانات الموجودة

في صيغة آسكي ASCII، كذلك تعد تكلفة تبني لغة الترميز الموسعة (XML) منخفضة نتيجة لارتباطها القوي بكل من لغة ترميز النصوص الفائقة (HTML)، واللغة المعيارية الموحدة الترميز (SGML).

وتجدر الإشارة إلى أن الشكل رقم (١١-١) ليس مقياساً محدداً، وما الأبعاد التي يتضمنها إلا مجرد أبعاد تصورية، تبين المبدأ الأساسي المتمثل في أن تكلفة تبني تقنية جديدة تعد أحد العوامل التي يجب أن تؤخذ بعين الاعتبار في جميع عمليات التشغيل المتداخل. كما أن العامل التقني يجب ألا ينظر إليه بمعزل عن العوامل الأخرى، وبدون وضع العامل التنظيمي في الاعتبار. فعندما يرغب منشئو إحدى المكتبات الرقمية في أن تعمل على نحو متبادل مع مكتبات أخرى، فإنهم غالباً يواجهون قضية المفاضلة بين البدائل المتاحة، ومن ثم اختيار أفضل الطرق المناسبة لمجتمعهم بشكل خاص، واختيار المعايير المقبولة بشكل جيد عموماً، والتي توفر لهم قدراً أقل من حيث الأداء الوظيفي. ولعل عملية اختيار الإصدارات الحديثة من البرمجيات تظهر مثل هذا التوتر، ففي الغالب توفر إحدى الإصدارات الجديدة للمكتبة الرقمية مزيداً من القدرة على أدائها الوظيفي، ولكن قلة من المستفيدين هم الذين يمكنهم الوصول إلى هذه المكتبات؛ حيث يستطيع منشئ أحد مواقع الويب - على سبيل المثال - أن يستخدم أكثر التيجان الأساسية للغة ترميز النصوص الفائقة، وأقل الصيغ الثابتة، والخدمات التي توفرها جميع إصدارات بروتوكول نقل الملفات الفائقة، في إنشاء هذا الموقع البسيط الذي يمكن لأي متصفح في العالم الوصول إليه ودعمه. وبدلاً من ذلك يمكن للمنشئ أن يختار أحدث نسخة من تقنيات الويب مع برامج تطبيقات جافا (جافا أبليت)، وصيغ

لغة ترميز النصوص الفائقة، وتجهيزات الأمن الداخلية، ونماذج عرض الصفحات (sheets)، وأدوات تحميل الصوت والصورة، وهي جميعها ستقدم أفضل خدمة فقط للمستفيدين الذين تتوافر لهم شبكات عالية السرعة مع أحدث المتصفحات، ومع ذلك فقد يرى آخرون أن جل هذه الخصائص غير مجدية بالنسبة لهم.

برامج بحث الويب :

تعد برامج الويب من أكثر النظم استخداماً لأغراض البحث الموزع أو اللامركزي، ومن أمثلة هذه البرامج إنفوسيك Infoseek، وليكوس Lycos، وألتافيستا AltaVista، وإكزايت Excite، وهي جمعيها نظم آلية تقوم بعملية تكشيف للمواد المتاحة على الإنترنت. وكما يظهر من الشكل رقم (١١-١) فإنها توفر مستوى متوسطًا من القدرة على الأداء الوظيفي، مع بعض الصعوبات القليلة في استخدامها والتي تتمثل في أمرين: أولهما أن مواقع الويب لا تتخذ أي إجراء خاص ليتم تكشيفها من قبل هذه البرامج، وثانيهما أنه ليس ثمة تكلفة يتحملها المستفيد سوى تكلفة الوقت الممل الذي يتحمله في مشاهدة الإعلانات التي تنشر عبر صفحاتها، وليس ثمة من سبيل لجعل برامج بحث الويب شائعة الاستخدام إلى أكبر نطاق سوى الجمع بين رفع كفاءة الأداء الوظيفي، والتحرر من صعوبات الاستخدام.

ومن الملاحظ أن الغالبية العظمى من هذه البرامج تشترك في بنيتها الفنية الأساسية، وإن كانت ثمة اختلافات كثيرة بينها في بعض التفاصيل، ولا يستثنى من ذلك سوى برنامج "ياهو"، الذي يعتمد في تنظيمه إلى عملية التصنيف الموضوعي. أما النظم الأخرى فتتكون من جزأين رئيسين: يعرف

أولهما بزاحف الويب أو الويب كرولر Web Crawler، (وهو الذي يقوم ببناء كشاف للمواد المتاحة على الإنترنت)، ويعرف ثانيهما بمحرك الاسترجاع research engine، (وهو الذي يسمح للمستفيدين بالبحث في هذا الكشاف).

زواحف الويب (ويب كرولرز) :

زاحف الويب هو برنامج تكشيف يتتبع الروابط الفائقة باستمرار، ويقوم بتجميع قائمة بالصفحات التي يعثر عليها، ويقوم كذلك ببناء كشاف مستمر النمو من صفحات الويب، وذلك من خلال تكرار الخطوات الأساسية القليلة التي يقوم بها. ويحتفظ الزاحف كذلك داخله بقائمة بالمحددات الموحدة لمواقع المصادر URLs المعروفة للنظام، سواء أكان قد تم تكشيف الصفحات المقابلة لهذه المحددات أم لم يتم تكشيفها بعد، ثم يقوم بعد ذلك باختيار المحدد الموحد الخاص بأي صفحة من صفحات الويب المرمزة عن طريق لغة ترميز النصوص الفائقة والتي لم يقم بتكشيفها من قبل، على أن يقوم باسترجاع هذه الصفحة ويعيدها مرة ثانية إلى النظام الآلي المركزي من أجل تحليلها، حيث يقوم هناك أحد برامج التكشيف الآلي بفحص الصفحة وإنشاء تسجيلة تكشيف لها، تضاف بعد ذلك إلى الكشاف الشامل، كما يقوم بالتقاط الروابط الفائقة الموجودة في هذه الصفحة والتي تشير إلى صفحات أخرى، بحيث يتم إضافة الروابط الفائقة الجديدة إلى قائمة المحددات الموحدة الموجودة لديه، وذلك من أجل استكشافها مستقبلاً أو في مرات لاحقة.

وتجدر الإشارة إلى أن هناك كثيراً من الاختلافات والمشكلات الفنية الدقيقة التي تكمن وراء هذا الإطار المبسط الذي تم تصويره، وتتمثل إحدى

التساؤلات عن المحدد الموحد الذي يجب على الزاحف أن يزوره بعد ذلك، ففي أي لحظة قد يكون هناك لدى زاحف الويب ملايين المحددات الموحدة التي يتم استكشافها وغير المتوقعة، لكن لا تتوافر له سوى معلومات قليلة يمكن الاسترشاد بها في اختيار المحدد التالي. وتشتمل معايير الاختيار على عدة تساؤلات أو عوامل منها: ما مدى حداثة هذا المحدد ؟ وما عدد المحددات الأخرى المرتبطة به ؟ وهل هذا المحدد يمثل إحدى الصفحات الرئيسية أم إحدى الصفحات الفرعية داخل مجموعة هرمية من الصفحات ؟

وبالرغم من كل ذلك، فلا تزال المشكلة الكبرى هي عملية التكشيف، فالزواحف تعتمد أساساً على عملية التكشيف الآلي - التي نوقشت في الفصل العاشر - في إنشائها للتسجيلات التي تقدم للمستفيدين، وبذلك فإن هذه الزواحف تواجه قضايا التكشيف الآلي الأساسية، فهناك ملايين من الصفحات التي يتم إنشاؤها من قبل أشخاص ذوي تصورات متباينة عن كيفية بناء المعلومات. كما أن الصفحات التي تعد مثالية لا تقدم إلا القليل من المؤشرات اللازمة لأغراض التكشيف الآلي، بل إن بعض منشئي الصفحات وبعض الناشرين يتعمدون التضليل؛ فيملؤون صفحاتهم بالمصطلحات التي يرجحون أن يستخدمها المستفيدون أملاً في أن تحظى صفحاتهم بالرتب الأعلى في سلم ترتيب قوائم الصفحات، والذي عادة ما يبنى على استفسارات البحث الشائعة. وأخيراً يمكن القول إنه بدون صفحات مبنية بناءً محكماً، أو ما وراء بيانات منتظمة، لا يمكن للتسجيلات الكشفية أن تحقق أعلى درجات الدقة عند عمليات البحث، بيد أن هذه التسجيلات يمكن أن تكون مناسبة لعمليات الاسترجاع المبسطة.

البحث في أحد كشافات صفحات الويب :

تسمح برامج بحث الويب للمستفيدين بالبحث في كشافاتها من خلال استخدام طرق استرجاع المعلومات التي أوضحناها في الفصل العاشر. وعادة ما تكون هذه الكشافات المجهزة بأسلوب يكفل لأعداد كبيرة من المستفيدين إمكانية البحث فيها بكفاءة وفي آن واحد. ونظراً لأن التسجيلات التي تتضمنها الكشافات قد لا تتمتع بالقدر الكافي من الدقة، بالإضافة إلى أن المستفيدين قد يفتقرون إلى التدريب الكافي للتعامل مع هذه السجلات، أو مع أساليب البحث، فإن برامج البحث تقوم باتباع استراتيجية تحديد جميع التسجيلات التي تضاهي الاستفسار، حتى وإن كانت هذه المضاهاة غير دقيقة، ثم تقوم بعرضها للمستفيد مرتبة طبقياً إلى حد ما.

ولعل معظم مستخدمي برامج الويب يتفقون على أنها برامج جيدة إلا أن ثمة كثير من الصعوبات الواضحة التي تكتنف استخدامها، فخوارزميات الترتيب الطبقي تفتقر إلى المعلومات الكافية التي يجب أن تبني عليها قرارات الترتيب. ونتيجة لذلك، فإن هذه البرامج قد تعطي بعض الصفحات الهامشية رتباً عالية لا تستحقها، ومن ثم ترد المواد الهامة في ذيل القائمة، وترد المواد الأقل أهمية أو التافهة على رأس القائمة، يضاف إلى ذلك أن برامج التكشيف يصعب عليها التعرف إلى المواد المكررة، مع أنها تحاول القيام بتجميع المواد المتشابهة، وبما أن هذه المواد المتشابهة يتوقع أن ترد جميعها في رتبة واحدة. فإن برامج التكشيف غالباً ما تقوم بإعداد قوائم مطولة من المواد المتماثلة تقريباً. ومن إحدى الطرق الجيدة للترتيب الطبقي

تلك الطريقة التي يتبعها محرك البحث جوجل Google، والتي تتمثل في قيامه بعمليات إحصائية للروابط. ويتميز جوجل بفاعلية جيدة في إيجاد المواد التمهيدية أو العامة حول الموضوع الذي يبحث عنه.

وعلى الرغم من أن زواحف الويب تقوم باستكشاف الويب بشكل مستمر بحيث يتم لها في النهاية العثور على كل شيء تقريباً، يلاحظ أن المواد الهامة قد لا يتم تكشيفها إلا بعد مضي عدة شهور من تحميلها على الويب، وعلى العكس من ذلك، يلاحظ أن البرامج لا تقوم بأداء وظيفتها على نحو جيد، والتي تتمثل في العودة مرة أخرى للتثبت من مدى سحب المواد أو استبعادها من على الويب، ولذلك نجد كثيراً من مداخل الكشافات تشير إلى مواد لم يعد لها وجود على الشبكة، أو أنها انتقلت من مواقعها إلى مواقع أخرى.

وهناك تهديد آخر يحيط بمدى كفاءة عملية تكشيف الويب يتمثل في أن زاحف الويب لا تستطيع تكشيف المواد التي لا يمكن الوصول إليها بشكل مباشر، فإذا كانت إحدى صفحات الويب خاضعة للحماية عن طريق إحدى صور المصادقة أو ضمانات الثقة authentication، أو إذا كانت إحدى صفحات الويب عبارة عن واجهة تفاعل لقاعدة بيانات، أو لمجموعات إحدى المكتبات الرقمية، فإن برامج التكشيف لا تتمكن من معرفة أي شيء عن المصادر الموجودة وراء الواجهة. وبما أن هناك الكثير والكثير من صفحات الويب عبارة عن واجهات تفاعل تتحكم فيها برامج جافا أو غيرها من برامج النصوص الأخرى، فإن برامج التكشيف تفقد كثيراً من المعلومات الهامة.

ومع أن نقاط الضعف هذه تبدو جوهرية، فإنه لا ينبغي المبالغة في ذلك، والدليل على ذلك يظهر من خلال الممارسة العملية؛ حيث يمكن لمستخدمي الويب الذين يتمتعون بقدر قليل من الخبرة، والمعتمدين في استخدامهم للويب على مجموعة من الأدوات المتنوعة، والتي تكون في الغالب خدمات بحث الويب -أن يصلوا للمعلومات التي يبحثون عنها. ومع أن هذه البرامج ليست مبرأة من كل نقص، فإنها تعد جيدة إلى حد بعيد، والأهم من ذلك أن استخدامها متاح مجاناً.

وتجدر الإشارة إلى أن معظم خدمات البحث كان لها جذورها في ما يعرف بمجموعات البحث research groups، إلا أنها ما فتأت أن أصبحت - بسرعة -شركات تجارية، ولقد كان لعدم فرض رسوم مالية على الخدمات الأساسية أثر كبير على كل من الإنترنت من ناحية، وعلى الشركات نفسها من ناحية أخرى؛ فقد أدى سعي هذه الشركات إلى تحصيل إيرادات مالية إلى التنافس المحموم في جلب الإعلانات، كما أن هذه الشركات اتجهت بسرعة إلى دخول الأسواق عبر وسائل معينة مثل منح تراخيص برمجياتها لشركات أخرى ترغب في بناء كشافات لمواقعها على الويب.

أما الجانب غير المرغوب في هذا النموذج التجاري فيتمثل في أنه يحد من الدافع إلى إنشاء كشاف شامل، فقد كانت برامج التكشيف تهدف في البداية إلى تكشيف الويب تكشيفاً شاملاً، ولكن في ظل النمو الذي تشهده الويب، وبحكم أن إدارة برامج البحث تحولت إلى مشروعات تجارية، فقد أصبحت الشمولية المنشودة في مرتبة ثانوية تالية للتحسينات في واجهات التفاعل وفي الخدمات المساعدة. ولا شك أن إنشاء كشاف جيد أو دقيق

ومحدث بمعنى الكلمة، يتطلب استثمارات ضخمة، وتقتنع معظم الشركات بأداء وظيفتها على نحو معقول، ولكن إذا زادت دوافعها، فسوف تكون كشافاتها أفضل مما هي عليه الآن.

اللوحة رقم (١١-١)

رُتب الصفحات وجوجل **Page Ranks and Google**

يعد تحليل الاستشهادات المرجعية من الأدوات الشائعة الاستخدام في العلوم، فالمقالات التي تستشهد بمقالات أخرى عادة ما تكون جميعها مرتبطة بعضها ببعض ارتباطاً موضوعياً. كما أن المقالات التي يكثر الاستشهاد بها عادة ما تكون أكثر أهمية من تلك المقالات التي لا يستشهد بها على الإطلاق. وقد قام كل من لورانس بيدج Lawrence Page، وسيرجي براين Sergy Brain وزملاؤهما في جامعة ستانفورد، بتطبيق فكرة هذا الأسلوب على الويب، مستخدمين الأنماط التي تتخذها الروابط الفائقة فيما بين الصفحات كأساس للترتيب الطبقي للصفحات، وطوروا برنامجاً تجريبياً لبحث الويب، يعرف باسم "جوجل".

فتخيل - على سبيل المثال - أنه باستخدام برامج بحث مختلفة للبحث عن كل ما يتصل "بجامعة ستانفورد"، وتبين أن ثمة أكثر من ٢٠٠,٠٠٠ صفحة تتصل بالتساؤل عن كلمة "ستانفورد Stanford"، ويلاحظ أن معظم برامج البحث يصعب عليها التمييز بين الصفحات ذات الطبيعة المحلية أو ذات القيمة الهامشية، وتلك الصفحات ذات الأهمية الكبرى فجميع برامج بحث الويب تُوجد أعدادًا هائلة من الصفحات التي تضاهي ذلك الاستفسار؛ أي ما يتصل بجامعة ستانفورد. ولكنها في معظم الحالات تخفق في ترتيب هذه الصفحات

ترتيباً طبقياً يعتبره المستفيدون دقيقاً بالشكل الذي يكفل للصفحات المهمة أن ترد في رأس القائمة.

وعندما قدم التساؤل عن جامعة ستانفورد لنظام جوجل لبحث الويب جاءت النتائج العشرة الأولى على النحو التالي:

- Stanford University Home page (www.stanford.edu)
- Stanford University Medical center (www.med.stanford.edu)
- Stanford University Library and information Resources (www.sul.stanford.edu/)
- Stanford Law school (www.leland.stanford.edu/group/law/)
- Stanford Graduate school of Business (www.gsb.stanford.edu/)
- Stanford University school of earth science (pangea.stanford.edu/)
- SUL: copyright & fair use (fairuse.stanford.edu/)
- Computer Graphic at Stanford University (www.graphic.stanford.edu)
- SUMMIT (Stanford University) Home page (summit. stanford.edu).
- Stanford medical Informatics (camis.standord.edu/).

وقد اتفقت الغالبية العظمى من المستفيدين على أن هذه القائمة جيدة.

والطريقة الأساسية التي استخدمها نظام "جوجل " لترتيب هذه الصفحات طريقة مبسطة، تتمثل في اعتمادها على قاعدة مفادها "أن الصفحات التي يتواتر الإشارة إليها من قبل الصفحات في شكل روابط فائقة تشير إلى صفحات أخرى، ترد في رتبة أعلى من الصفحات التي تقل

الإشارة إليها، كما أن الروابط التي وجدت في الصفحات التي احتلت رتباً أعلى تأخذ وزناً أكبر من الروابط التي وجدت في الصفحات التي احتلت رتباً متدنية. وبما أن صفحات الويب حول العالم تقوم بعمل روابط إلى الصفحة الرئيسية لكلية الحقوق بجامعة ستانفورد، فإن هذه الصفحة تقوم بدورها بعمل روابط إلى العديد من الصفحات الأخرى، مثل الصفحة الرئيسية للجامعة، ومن ثم فإن هذه الأخيرة جاءت في رتبة متقدمة بسبب الإشارة إليها من قبل صفحة احتلت رتبة متقدمة كذلك.

ويتطلب حساب رتب الصفحات عملية حسابية ذكية، ولكي تفهم الفكرة الأساسية لهذه العملية، تخيل أن أمامك مصفوفة كبيرة تضم كل صفحة موجودة على الإنترنت، وأمام كل صفحة قائمة بالروابط التي تشير إليها، ففي بداية الأمر، يتم ترتيب الصفحات على نحو متساوٍ، ثم يتم بعد ذلك حساب عملية الترتيب الجديدة لهذه الصفحات على أساس عدد الروابط الموجودة أمام كل صفحة، على أن يتم عمل وزن لكل صفحة بناء على رتبة الصفحات التي وردت بها الروابط لتلك الصفحة linking pages، منسوبة إلى عدد الروابط المأخوذة من كل صفحة. ثم يتم بعد ذلك استخدام هذه الرتب لتكرار آخر، وتستمر العملية على هذا النحو حتى تنتهي العملية الحسابية.

وتعد العملية الحسابية الآلية computation الحقيقية تحسينًا لهذه الطريقة، ففي عام ١٩٨٨م كان جوجل يضم ما يقرب من ٢٥ مليون صفحة، تم اختيارها عن طريق عملية تم أخذها من رتب الصفحات التي أشارت أو قدمت ربطاً لهذه الصفحات. وتتوافر للبرنامج عوامل وزن لحساب الصفحات التي

ليس لها روابط، ولحساب مجموعات الصفحات التي بها روابط لصفحة واحدة أخرى فقط. ويرفض البرنامج الصفحات التي يتم إنشاؤها بصورة آلية أو ديناميكية من قبل نصوص واجهات بوابات العبور الموحدة Common Gateway Interface (CGI). ذلك لأن النظام كان قادراً على جمع هذه الصفحات وتكشيفها وترتيبها في خمسة أيام باستخدام حاسبات محطات العمل القياسية فقط، ولكن ليبرهن على قوة حاسبات اليوم.

إن استخدام الروابط لإنشاء رتب الصفحات يساعد على حل مشكلتين من المشكلات التي تضعف عمل برامج بحث الويب، وهاتان المشكلتان تتمثلان في التساؤلين التاليين: بما أن البرامج لا تستطيع تكشيف كل صفحة متاحة على الويب في الوقت نفسه الذي تظهر فيه، فما الصفحات التي يجب أن تكشف قبل غيرها؟ وكيف تتمكن البرامج من ترتيب الصفحات التي تسترجع عند طرح التساؤلات البسيطة حتى يتم إعطاء الأولوية للصفحات الأكثر أهمية ؟

المكتبات الرقمية الاتحادية :

إن تخفيف حدة الجدل القائم حول قضيتي الأداء الوظيفي وتكلفة تبني تقنية معينة يعتمد على الظروف التي يتم في سياقها هذا الجدل، فقد يكون من المناسب في بعض الأحيان اختيار تقنية بسيطة، والسعي الحثيث لتحقيق معدلات تغطية واسعة غير متعمقة من التشغيل المتداخل، وقد يكون من الحكمة في أحيان أخرى أن يتم اختيار تقنية تتمتع بقدر عال من الكفاءة في الأداء الوظيفي، ولكن ذلك يتطلب تحمل تكاليف عالية، ولا يقدر على تبني

هذه الأساليب المكلفة سوى المكتبات الرقمية ذات الطموحات الكبيرة، وإن كانت سوف تحقق مستوى عاليًا في أدائها الوظيفي.

ومصطلح المكتبة الرقمية الاتحادية يستخدم ليشير إلى مجموعة من المنظمات التي تعمل بعضها مع بعض بشكل رسمي أو غير رسمي، متفقة فيما بينها على دعم مجموعة من الخدمات والمعايير المشتركة، ومن ثم توفير إمكانية التشغيل المتداخل بين أعضائها. وتجدر الإشارة إلى أن أعضاء هذا الاتحاد يمكن أن يكون لكل منهم نظام يختلف عن النظم الأخرى اختلافاً تاماً، إلا أن ذلك لا يجسد مشكلة ما دامت جميعها تدعم مجموعة من الخدمات المتفق عليها. ويتطلب الأمر اتفاقهم على كل من المعايير الفنية والسياسات، بما في ذلك الاتفاقيات المالية وتشريعات الملكية الفكرية، وسياسات الأمن والخصوصية.

ويقدم بحث أجري في جامعة إلينويز في أربانا شامبين University of Illinois at Urbana Champaign مثالاً للصعوبات التي تكتنف عملية التشغيل المتداخل، ففي الفترة من ١٩٩٤- ١٩٩٨م وكجزء من مبادرة المكتبات الرقمية، بدأ فريق من العاملين في مكتبة جرينجر الهندسية Grainger Engineering Library بإنشاء مكتبة اتحادية لمقالات الدوريات المنشورة من قبل العديد من الناشرين العلميين البارزين. وبما أن كل ناشر من هؤلاء كان قد خطط لإتاحة دورياته بصيغة اللغة المعيارية الموحدة للترميز (SGML)، فقد بدا ذلك وكأنه فرصة لإنشاء ذلكم الاتحاد. ولأنه كان من المفترض على الجامعة أن تقدم خدمات مركزية، كخدمات البحث، فقد كان على الناشرين تأمين تلك الدوريات. لكن عدم تناغم الأساليب التي استخدم بها الناشرون

اللغة المعيارية الموحدة للترميز؛ قد سبب مشكلات في هذا الصدد، حيث كان لكل ناشر من هؤلاء الناشرين تعريفه الخاص بشكل الوثيقة. وقد اضطرت الجامعة إلى القيام بمهام مطولة للتوفيق بين الدلالات اللفظية لمعرفات أنواع الوثائق the semantics DTDs، من أجل إتاحة إمكانية اشتقاق المعلومات الكشفية، ومن أجل تطوير واجهة متماسكة للمستخدمين. وقد ثبت أن ذلك كان عملاً مضنياً مما حدا بالجامعة أن تلجأ إلى استنساخ المعلومات من حاسبات الناشرين وتحميلها على نظام واحد، ثم تحويلها إلى معرف واحد من معرفات أنواع الوثائق. وإذا كان قد قدر لمجموعة بحث في جامعة مرموقة أن تواجه هذه الصعوبات مع جزء متناغم نسبياً من المعلومات، فليس من المدهش أن يواجه الآخرون مثل تلك الصعوبات.

اللوحة رقم (١١-٢)

مكتبة جامعة إلينويز الاتحادية للإنتاج الفكري في العلوم

The University of Illinois Federated Library of Scientific Literature

تعد مكتبة جرنجر الهندسية (The Grainger Engineering Library) في جامعة إلينويز مركزاً لأحد النماذج الأولية prototype لمكتبة اتحادية لمقالات الدوريات العلمية. وقد بدأ العمل فيها كجزء من مبادرة المكتبات الرقمية بقيادة كل من بروس تشاتز Brauce Schatz ووليم ميسكو William Mischo. وبحلول عام ١٩٩٨م وصلت مجموعة المقتنيات التي خضعت للاختيار إلى ٥٠,٠٠٠ مقالة من المقالات المنشورة في دوريات كل من معهد المهندسين الكهربائيين والإلكترونيين (IEEE)، وجمعية الحاسب في هذا المعهد، والجمعية الفيزيائية الأمريكية، وجمعية المهندسين المدنيين

الأمريكيين، والمعهد الأمريكي لعلوم الطيران والفضاء. وقد اتفق على أن تقوم كل جهة من هذه الجهات بتقديم مقالات من دورياتها العلمية مهيأة باللغة المعيارية الموحدة للترميز في الوقت نفسه الذي تنشر فيه تلك الدوريات مطبوعة.

وقد طبق هذا النموذج الأولي مفاهيم استرجاع المعلومات من النص المرمز marked-up text والتي نوقشت كثيراً ولم تطبق إلا قليلاً. وقد بدأت المرحلة الأولى (التي اتسمت بالصعوبة والتعقيد) بعملية التوفيق بين معرفات أنواع الوثائق التي استخدمها ناشرو الدوريات، حيث كان كل منهم يستخدم معرفاً خاصاً به لعرض العناصر البنائية لوثائقه. وقد تجسدت بعض أوجه الخلاف في الجوانب الدلالات اللفظية، حيث كان يعبر عن عنصر المؤلف في بعض الأحيان - على سبيل المثال - بالتاج <author>، في حين يعبر عنه في أحياناً أخرى بالتاج <aut>، وفي أحيان ثالثة بالتاج <au>. وكانت الخلافات الأخرى تعكس أوجه اختلاف دلالية جوهرية. ومن أجل تنفيذ عمليتي التكشيف والاسترجاع، لجأ القائمون على المشروع إلى وضع برمجية تقوم بعمل تخطيط لتيجان كل معرف من معرفات نوع الوثائق على هيئة صيغة مقبولة من الجميع (canonical set) على أن تقوم واجهات التفاعل مع المكتبة الرقمية باستخدام هذه التيجان، بحيث يستطيع المستفيد أن يبحث عن نص في سياق معين، كما هو الحال مع دليل الصور أو الأشكال figure legend. وفي مخطط الشكل رقم (١١-١) يقع استخدام هذه المجموعة من التيجان في أعلى اليسار. ومع أن عملية تحويل الترميز التي أعدت لجميع مفردات هذه

المقتنيات إلى هذه الصيغة تتطلب تكلفة عالية كلما أضيفت مجموعة مقتنيات جديدة إلى الاتحاد، فإنه يهدف إلى توفير مستوى عال من الأداء الوظيفي.

وبسبب الصعوبات الفنية، تمثلت الخطوة الأولى من التنفيذ في تحميل جميع الوثائق في مستودع واحد في جامعة إلينويز، على أن تدعو الخطط المستقبلية إلى استخدام المستودعات التي تم إعدادها من قبل الناشرين. وَثَمَّ اهتمام آخر يقضي بتوسعة مجال المقتنيات لتشمل قواعد البيانات الببليوجرافية، وفهارس، وكشافات أخرى.

وبالرغم من أنه قد ثبت أن عملية التنفيذ الأولى لم تكن أرضاً قوية لدراسة المستفيدين ورغباتهم، فإن توفيرها لأساليب إضافية قوية لإجراء عمليات من جانب المستفيدين كان أمراً محموداً، ولكنه شجع على تلقي الاستفسارات. وقد أشار المستفيدون إلى أن الأشكال البيانية والمعادلات الرياضية غالباً ما تكون موضحة بشكل كبير في محتوى المقالات أكثر من مستخلصات تلك المقالات والفقرات الختامية الخاصة بالنتائج التي تتضمنها تلك المقالات. كما أظهرت هذه التجارب - مرة أخرى - أن المستفيدين يواجهون صعوبات كبيرة في العثور على الكلمات الصحيحة التي يجب استعمالها في استراتيجيات البحث عندما لا يكون هناك ضبط للمصطلحات المستخدمة في البحوث وفي مستخلصاتها، وفي نظام البحث.

الفهارس المباشرة ومعيار Z39.50 :

لكثير من المكتبات فهارس مباشرة لمقتنياتها التي يتاح الوصول إليها مجاناً من خلال الإنترنت، ويمكن اعتبار هذه الفهارس شكلاً من أشكال

الاتحادات. وعادة ما تتبع قواعد فهرسة الأنجلو الأمريكية Anglo-American Cataloging Rules في إعداد تسجيلات هذه الفهارس من خلال استخدام صيغة الفهرسة المقروءة آلياً المعروفة بمارك، وتقوم المكتبات بتقاسم عملية إعداد هذه التسجيلات فيما بينها بهدف تخفيض التكلفة. وقد قام المكتبيون بتطوير معيار Z39.50 (الذي سيرد وصفه في اللوحة رقم ١١-٣)، من أجل تلبية احتياجات ما يتصل بكل من تقاسم إعداد التسجيلات، وعمليات البحث اللامركزي. وفي الولايات المتحدة نشطت كل من مكتبة الكونجرس ومركز الحاسب الآلي للمكتبات على الخط المباشر OCLC، ومجموعة مكتبات البحث في تطوير هذا المعيار ونشره، وقد كان هناك العديد من التطبيقات المستقلة لمعيار Z39.50 في المواقع الأكاديمية، والموردين التجاريين. كما تم استيعاب التكلفة العالية للانضمام لمثل هذا الاتحاد عبر العقود الماضية، وتم تعويض هذه التكاليف من خلال الوفورات المالية التي تحققت من جراء عمليات الفهرسة المشتركة.

وتتمثل إحدى جوانب التطبيقات الأساسية لمعيار Z39.50 في مجال الاتصالات بين الحاسبات الخادمة. ويمكن لنظام الفهرسة في إحدى المكتبات الكبرى أن يستخدم هذا المعيار لبحث مجموعة من الفهارس المتشابهة، ليرى ما إذا كان أي منها يوجد لديه نسخة من عمل ما، أو قامت بإعداد تسجيلة له. ويمكن للمستفيدين النهائيين أن يستخدموا حاسباً عميلاً واحداً يدعم معيار Z39.50 لبحث العديد من الفهارس، سواء تم ذلك على التوالي أو على التوازي. وقد جنت المكتبات وروادها فوائد جمة من خلال عمليات تقاسم الفهارس بهذه الطرق، وبالرغم من ذلك، لا تزال

عملية التشغيل المتداخل بين الفهارس المباشرة public access غير مكتملة الأركان. ومع أن بعض تطبيقات معيار Z39.50 تمتع ببعض المزايا التي لا تتمتع بها غيرها، فإن السبب الأساسي لعدم الاكتمال يتمثل في أن الفهارس الفردية يتم إعدادها من قبل أشخاص يعملون وفق احتياجات مجتمعاتهم المحلية، ولم يكن دعم المؤسسات الأخرى يحظى بالأولوية على الإطلاق. وحتى عندما تتقاسم المؤسسات نسخاً متوافقة من البروتوكول Z39.50، فإن الاختلافات في كيفية تنظيم الفهارس وإتاحتها للعالم الخارجي ستظل باقية.

اللوحة رقم (١١-٣)

معيار **Z39.50**

في رحاب مجتمع المكتبات تم تطوير المعيار المعروف بـ Z39.50، والذي يسمح لأحد الحاسبات (العميل) بالقيام بعمليات البحث والاسترجاع للمعلومات التي توجد لدى حاسب آخر (خادم قاعدة البيانات)، ويعد هذا المعيار هامًّا بحكم بنيته الفنية، وبسبب شيوع استخدامه في نظم المكتبات. ومع أن هذا البروتوكول لا يعد- من الناحية النظرية - مخصصاً للتعامل مع أشكال محددة من المعلومات، أو نوعية معينة من قواعد البيانات، فإن معظم توجهات تطويره تركزت على البيانات الببليوجرافية، كما تركزت معظم تطبيقاته على البحوث التي تستخدم السمات الببليوجرافية في بحث قواعد بيانات تسجيلات مارك، واسترجاع هذه التسجيلات إلى الحاسب العميل.

ويبنى المعيار Z39.50 على وجهة نظر مجردة للبحث في قواعد البيانات، حيث يفترض أن الحاسب الخادم يقوم باختزان مجموعة من قواعد البيانات مع كشافات قابلة للبحث. وتبنى التفاعلات على فكرة أو مفهوم

"الجلسة session"، إذ يقوم الحاسب العميل بالاتصال بالحاسب الخادم، ثم يقوم بإجراء سلسلة من التفاعلات، ثم يقوم بإنهاء الاتصال بعد ذلك. وأثناء وقت الجلسة يقوم كل من الخادم والعميل بتذكر حالة تفاعلاتهما، ومن المهم أن نعي أن العميل هنا هو جهاز حاسب آلي، وأن تطبيقات المستخدم النهائي للبروتوكول Z39.50 تستلزم واجهة للمستخدمين تستخدم لأغراض الاتصال بالمستفيد، ولا يقوم البروتوكول بطلب بيانات عن شكل واجهة المستفيد، ولا عن كيفية اتصالها بالحاسب العميل الذي يدعم بروتوكول Z39.50.

وتبدأ الجلسة الفعلية بقيام العميل بإجراء الاتصال بالخادم، ومن ثم تبادل المعلومات الأولية بينهما، مستخدماً في ذلك وسيلة "خاصية init"، ويقوم هذا التبادل الأولي بعمل اتفاقية على الأمور الأساسية، مثل الحجم المفضل للرسالة، كما يمكن أن تحتوي كذلك على إجراءات "المصادقة authentication " وإن كانت عمليات المصادقة الحقيقية تتم خارج نطاق المعيار. ويمكن للعميل بعد ذلك استخدام خدمة "التفسير explain" ليستفسر من الخادم عن كل من قواعد البيانات المتاحة للبحث، وعن الحقول القابلة للبحث فيها، وعن تركيب كلمات البحث وجمله syntax، وعن أشكال النص التي يدعمها، إلى غير ذلك من خيارات.

وتسمح "خدمة البحث search service" للعميل بأن يقدم استفسارًا إلى قاعدة البيانات كما في المثال التالي :

"في قاعدة البيانات المسماة "Books" ابحث عن جميع التسجيلات التي تشتمل حقول العنوان فيها على القيمة "Evangeline"، والتي تشتمل حقول بيانات التأليف فيها على القيمة "Longfellow".

ومع أن المعيار يقدم عدة بدائل لتركيب الكلمات وجمل البحث[١] التي تستخدم لتحديد مجال البحوث، فإن الاستفسارات البولينية[٢] هي التي يتم تنفيذها بشكل شائع، حيث يقوم بتنفيذ البحث، واسترجاع النتائج. ومن الإمكانيات المتميزة للبروتوكول Z39.50 تمكن الخادم من حفظ نتائج البحث، بحيث يمكن إحالة العميل في استفسار لاحق إلى هذه النتائج. وهكذا يستطيع العميل تكوين مجموعة ضخمة من النتائج الناجمة من خلال الاستفسارات الدقيقة، كما يمكنه طلب عرض أي تسجيلة من هذه المجموعة، بدون القيام بالبحث الكامل لقاعدة البيانات مرة أخرى.

وبناء على خصائص استفسارات البحث، يمكن أن تسترجع تسجيلة واحدة أو أكثر للعميل، كما يتيح المعيار عدة أساليب متنوعة للعملاء تمكنهم من معالجة نتائج البحث، مثل إمكانات الفرز والحذف، وعندما تنتهي عملية البحث، يحتمل أن يبدأ العميل بإرسال استفسار جديد. وهذه التساؤلات تجعل الخادم يقوم بإرسال تسجيلات محددة من مجموعة النتائج إلى العميل في شكل محدد. وتحتوي الخدمة الحالية على عدة خيارات تتصل بعملية التحكم

(١) تعرف باستراتيجيات البحث (المترجمان).

(٢) تعرف بالروابط البولينية أو المنطقية (المترجمان).

فـي المحـتوى وفـي الـصيغ، وتتـصل بعملـيات إدارة التـسجيلات الكثيـرة أو مجموعات النتائج الكبيرة.

وفـضلاً عـن خدماتـه الأساسـية، تتوافـر للبـروتوكول Z39.50 إمكانـيات تـصفح الكـشافات، وإمكانـيات ضـبط عملـيات الوصـول Access، وإمكانـيات إدارة المـوارد، كمـا أنـه يـدعم الخـدمات الموسـعة التـي تـسمح بمـزيد مـن التوسعات extensions. إنه حقًا لمعيار ضخم ومرن.

المكتـبة المـرجعية الفنـية الـشبكية لعلـوم الحاسـب NCSTRL، وداينيست Dienst :

"الفهرس الموحد" هو فهرس واحد يشتمل علـى تسجيلات مقتنيات عدة مكتبات، وتساعد هذه الفهارس الموحدة (التـي تستخدمها المكتبات منذ وقت طـويل قـبل ظهـور الحاسـبات الآلـية) علـى إجـراء عملـية الـبحث المـوزع اللامركزي، من خلال تجميع المعلومات التي سيجري بحثها في فهرس واحد. وتعد خدمـة البحث في الويب بمثابة فهارس موحدة للويب، حتى وإن تضمنت إحداها تسجيلات فهارس غير معدة بشكل صحيح. وثمة طريقة بديلة للبحث الموزع تتمثل في أن يكون لكل مجموعة مقتنيات كشافها القابل للبحث؛ بحيث يقوم برنامج البحث بإرسال الاستفسارات إلى هذه الكشافات المنفصلة، ثم دمج النتائج جميعها وتقديمها للمستفيد.

والمكتبة المرجعية الفنية الشبكية لعلوم الحاسب (NCSTRL) هي اتحاد لمجموعات المكتبات الرقمية الهامة للباحثين في مجال علوم الحاسب، وهي تستخدم بروتوكولاً يسمى داينست Dienst، وفي سبيل تخفيض تكاليف تقبله،

بنـي هـذا البـروتوكول علـى مجمـوعة متـنوعة مـن المعاييـر الفنـية المألـوفة لعلمـاء الحاسـب الـذين يعتبـرون مـن أكثـر مسـتخدمي نظـام اليونكس UNIX، والإنترنت، والويب. وتجدر الإشارة إلى أن الإصدارة الأولى من بروتوكول داينست كانت تقوم على إرسال الاستفسارات البحثية إلى جميع الحاسبات الخادمة، ونظراً لأن أعداد هذه الحاسبات تزايدت، فقد توقف العمل بهذا الأسلوب، وإذا لم يكن هناك خادم معين موجود، فإن النظام برمته يضعف. والآن يتوافر لداينست كشاف أصلي master index يعد شكلاً من أشكال الفهارس الموحدة.

اللوحة (١١–٤)

المكتبة المرجعية الفنية الشبكية لعلوم الحاسب، ونموذج داينست للبحث الموزع

NCSTRL and the Dienst Model of Distributed Searching

المكتبة المرجعية الفنية الشبكية لعلوم الحاسب هي مكتبة موزعة أو لامركزية للإنتاج الفكري في علوم الحاسب، وخاصة التقارير الفنية منها. وتقوم الجهات المتعاونة بتحميل مجموعاتها على حاسبات خادمة محلية، ويتم الاتصال أو الوصول إلى هذه الحاسبات الخادمة، إما عن طريق بروتوكول نقل الملفات FTP، أو عن طريق بروتوكول مسمى داينست Dienst، وهذا البروتوكول الأخير هو بروتوكول للمكتبة اللامركزية، وقد قام بتطويره كل من جيم ديفس Jim Davis من شركة زيروكس، وكارل لاجوز Carl lagoze من جامعة كورنيل، وتم تطوير هذا البروتوكول كجزء من مشروع التقارير الفنية لعلوم الحاسب CSTR project ، الذي أشير إليه في

الفصل الرابع. وفي بداية المشروع كانت هناك خمسة جامعات متعاونة، ازدادت لتصل إلى أكثر من مئة جهة من جميع أنحاء العالم بحلول عام ١٩٩٨م، كانت ثلاث وأربعون جامعة منها تشغل خادمات داينست. وتقوم المكتبة المرجعية الفنية الشبكية لعلوم الحاسب وداينست على الجمع بين الخدمات اليومية مع قاعدة اختبار test bed لأغراض إدارة المعلومات الموزعة. وتعد هذه المكتبة من الأمثلة القليلة لمجموعات البحث التي تقوم بتشغيل إحدى الخدمات العملية للمكتبات الرقمية.

وتقوم البنية الفنية لداينست على تقسيم خدمات المكتبة الرقمية إلى أربع فئات رئيسية هي: المستودعات والكشافات،والمجموعات، وواجهات المستفيدين. وتوفر بروتوكولاً مجانياً يعرف بهذه الخدمات؛ ويعمل هذا البروتوكول على دعم عمليات البحث الموزع للمجموعات التي تدار على نحو مستقل. ويتوافر لكل خادم كشاف للمواد التي تختزن فيه. ولكي تتم عملية البحث من خلال استخدام الإصدارات الأولية من داينست، كان برنامج واجهة المستفيدين يقوم بإرسال الاستفسار إلى جميع مواقع داينست بحثاً عن الكائنات objects التي تضاهي هذه التساؤلات، وكان على واجهات المستفيدين أن تنتظر حال وصول استجابات من الحاسبات الخادمة. وقد كان هذا هو البحث اللامركزي في شكله الأساسي، وقد بدت كثير من المعوقات بسبب الزيادة الكبيرة في أعداد الحاسبات الخادمة، وقد كانت المشكلة الأساسية هي أن جودة الخدمة التي يراها أحد المستفيدين كانت تعتمد على مستوى الخدمة التي كان يقدمها أسوأ مواقع داينست. ولفترة معينة ظل الخادم الموجود في جامعة

كارنيجي ميلون خارج حدود التعويل عليه، وإذا فشل في الاستجابة لتساؤل معين، كان على واجهة المستفيد الانتظار حتى تصدر الإشارة الآلية على انتهاء الوقت، بل إنه حتى عندما كانت جميع الحاسبات الخادمة في طور العمل، كانت تحدث في الغالب تأخيرات مطولة بسبب صعوبة الاتصال ببعض الحاسبات الخادمة القليلة.

ولا شك أن البحث البطيء يدعو إلى تذمر المستفيد، كما أن البحث يكون أكثر خطراً عندما يفقد بعض المجموعات؛ فالفشل في بحث جميع الكشافات يعني أن الباحث قد يفقد معلومات هامة، وما المشكلة ببساطة إلا بسبب الحاسب الخادم.

وقد تم إعادة تصميم داينست لتفادي هذه المشكلات، وتقسم الآن المكتبة المرجعية الفنية لعلوم الحاسب إلى مناطق، وقد كان هناك في البداية مركزان إقليميان في الولايات المتحدة، وأربعة في أوروبا. ومع هذا النموذج الإقليمي، يتم وضع الكشاف الأصلي في مكان مركزي (جامعة كورنيل)، كما يتم اختزان نسخ قابلة للبحث من هذا الكشاف في المراكز الإقليمية، بحيث يكون هناك في كل موقع إقليمي كل شيء يريد المستفيد البحث عنه، وأي معلومات يريدها. ولا يتصل المستفيد بالمواقع الفردية للمجموعات إلا من أجل استرجاع المواد المختزنة بها. وتترك مسألة اختيار أي المواقع الإقليمية التي يتم الاتصال أو البحث فيها بها للمستفيد، حيث يمكن للمستفيد، نتيجة لمشكلات الاتصال بالإنترنت، أن يحصل على أفضل الأداء من أحد المواقع الإقليمية البعيدة من تلك المواقع القريبة له من الناحية الجغرافية.

اتجاهات البحث في السبل البديلة للبحث الموزع :

إن النجاح أو الفشل الذي يحدث في سياق أحد التجمعات الاتحادية federationيعد تحدياً تنظيمياً أكثر منه تحدياً فنياً. ومن المؤكد أن بعض أعضاء هذا الاتحاد يقدم خدمات تفوق ما يقدمه الآخرون، كما أن مستويات الدعم التي يقدمها كل منهم تختلف اختلافاً كبيراً، وبالرغم من ذلك، لا ينبغي ربط جودة الخدمة بأسوأ المنظمات أداءً لوظائفها. وتصور اللوحة رقم (١١-٤) كيف أعيد تصميم نظام داينست، الذي تستخدمه المكتبة المرجعية الفنية الشبكية لعلوم الحاسب، لكي يلبي تلك الحاجة.

وتضع كل خدمة من خدمات المعلومات بعض الافتراضات الضمنية عن السيناريوهات التي تدعمها، والاستفسارات التي تقبلها، وأنواع الإجابات التي توفرها. وهذه تنفذ على أنها من قبيل التيسير، كما هو الحال بالنسبة لخدمات البحث عن المعلومات وغيرها من خدمات التصفح والتنقية والاشتقاق extract؛ فالمستفيد عادة ما يطمح إلى الحصول على معلومات متماسكة تلبي احتياجاته الشخصية، مع أن مصادر المعلومات في هذا العالم مترامي الأطراف والتي تعد متماسكة - تتفاوت فيما بينها، فكيف تستطيع إذن مجموعة متنوعة من المنظمات أن توفر للمستفيدين الحاسبات الخادمة التي تم تصميمها وفقاً لسيناريوهات ضمنية مختلفة أن تقدم استكشافاً فعالاً لمصادرها بدون عملية توحيد صارمة؟ كما أن هناك تساؤلات أو عقبات فنية صعبة تواجه إحدى الخدمات الفردية التي تدار على نحو مركزي، وقد باتت هذه القضايا معقدة فعلاً عندما خضعت مصادر المعلومات لسيطرة جهات لها استقلاليتها.

إن التحديات التنظيمية تعد بالغة الصعوبة لدرجة أنها تحد من الخيارات الفنية المتاحة، وباستثناء ما يحدث في الاتحادات القوية، فإن الأمل الوحيد في التقدم يكمن في إيجاد أُطر عمل فنية، بالتوازي مع استراتيجيات تقبل هذه الأطر، يمكن للمنظمات أن تقبلها بصورة متدرجة. وبالنسبة لكل طريقة من هذه الطرق يجب أن يكون هناك بديل ذو مستوى أدنى (عادة ما يكون هو الحالة الموجودة حالياً status quo)، وذلك حتى لا تعاق الخدمات عن أداء أي شيء جدير بالاهتمام بسبب قلة من النظم. ولذلك، فإن هناك في المكتبة المرجعية الفنية الشبكية لعلوم الحاسب - على الرغم من أن داينست يعد البروتوكول المفضل - أكثر من نصف المواقع تحمل مجموعاتها على حاسبات خادمة يدعمها بروتوكول نقل الملفات ftp.

إن كثيراً من البحوث في مجال البحث الموزع تبدأ ببناء فهارس موحدة مما وراء البيانات التي يقدمها المنشئ أو الناشر. وقد كان هذا أحد الدوافع الكامنة وراء معيار دبلن كور؛ حيث تحتاج النظم الآلية إلى جمع ما وراء البيانات هذه وضمها في كشاف قابل للبحث.

وهناك مجال آخر من مجالات البحث في هذا المجال يهدف إلى تطوير أساليب تعمل على تقييد عمليات البحث على أكثر المجموعات نجاحاً، فالمستفيدون نادراً ما يرغبون في البحث في كل مصدر من مصادر المعلومات المتاحة على الإنترنت، ولكنهم يرغبون في البحث في فئات محددة، مثل فهارس المنفردات، أو كشافات الإنتاج الفكري الطبي. ولذلك فإن هناك حاجة إلى بعض الوسائل التي تمكن المجموعات من تقديم ملخصات

لمحتوياتها. مع ملاحظة أن هذا يكتسب أهمية خاصة عندما يكون الوصول مقيداً بضابطي المصادقة وآليات دفع الرسوم. وإذا تم توفير الوصول المجاني إلى أحد المصادر، فإن برنامجاً خارجياً يمكن - على الأقل من الناحية النظرية - أن يقوم بعمل ملف إحصائي بأنواع المواد وبالمصطلحات المستخدمة. وعندما يتاح لمستفيد خارجي إمكانية الوصول فقط من خلال إحدى واجهات البحث، فإن مثل هذا التحليل لا يكون ممكناً.

وفي عام ١٩٩٦م، قام لويس جرافانو Luis Gravano بجامعة ستانفورد- بدراسة الكيفية التي يمكن بها أن يقوم أحد الحاسبات العميلة بدمج النتائج المتحصل عليها من خدمات بحث مستقلة، وقام هذا الباحث بتطوير بروتوكول مخصص لهذا الغرض عرف باسم ستارتز STARTS. وكان هذا البروتوكول مشروعًا مشتركًا بين جامعة ستانفورد من ناحية، وعدد من شركات الإنترنت الرائدة من ناحية أخرى. ويدل مدى الترحيب الذي أبدته تلك الشركات للانضمام إلى جهود هذا المشروع على أنها تدرك أن أفكاره تكتسي أهمية خاصة وأساسية بالنسبة للبحث الموسع عبر الإنترنت، وبقليل من جهود التوحيد يمكن أن تؤدي إلى تطور كبير في عمليات البحث.

وقد كان جرافانو في تحليله ينظر إلى المعلومات المتاحة عبر الإنترنت على أنها مجموعة كبيرة من المواد، وكل منها منظم على نحو مختلف، كما أن لكل منها محرك بحث خاصًّا بها. والفكرة الأساسية تتمثل في تمكين الحاسبات العميلة من استكشاف الخصائص العامة لمحركات البحث وللمجموعات التي تحتفظ بها. غير أن التحدي يكمن في اختلاف محركات

البحث، كما أن خصائص هذه المجموعات هي الأخرى مختلفة. والصعوبة ببساطة ليست في أن لواجهات التعامل أساليب مختلفة في صياغة استراتيجيات البحث فيها، بل في أن صياغة الاستراتيجية الواحدة يجب أن تعاد بعدة طرق لتتلاءم مع النظم المختلفة، كذلك فإن الخوارزميات المستخدمة هي الأخرى مختلفة تماماً، فبعض النظم تستخدم الطرق البولينية أو المنطقية في البحث، وبعضها الآخر له أساليبه الأخرى في ترتيب النتائج. كما أن محركات البحث التي تقوم بتقديم قائمة مرتبة طبقياً بالنتائج تعطى مؤشرات قليلة عن الكيفية التي تم بها ترتيب هذه النتائج. وغالباً ما يكون حساب عملية الترتيب الطبقي سراً تجارياً. ونتيجة لذلك يكون من غير الممكن دمج قوائم النتائج الطبقية التي ترد من مصادر متعددة في قائمة واحدة وبترتيب طبقي مقبول؛ ذلك لأن الترتيب الطبقي يتأثر تأثراً كبيراً بطبيعة الكلمات المستخدمة في المجموعة، ولذلك فإن دمج النتائج المتحصلة من مصدرين يستخدمان خوارزمية الترتيب الطبقي نفسها سيكون محفوفاً بالمصاعب. ويعمل بروتوكول ستارتز على تمكين محركات البحث من عمل تقارير عن خصائص مجموعاتها، وأساليب الترتيب التي تتبعها، حتى يستطيع برنامج أحد الحاسبات العميلة أن يدمج النتائج المتحصل عليها من مصادر متعددة.

اللوحة رقم ١١-٥

البنية الفنية لمشروع هارفست **The Harvest Architecture**

كان هارفست مشروعاً بحثياً عن أساليب البحث الموزع، تبناه مايكل شوارتز Michael Schwartz، ثم انتقلت رعايته إلى جامعة كلورادو، وعلى الرغم من أن هذا المشروع انتهى عام ١٩٩٦م، إلا أن الأفكار الأساسية التي

ارتبطت بتطوير بنيته الفنية لا تزال مناسبة لهذا المجال. وتركز الفكرة الأساسية لهذا المشروع على تقسيم الوظائف الأساسية في نظام البحث المركزي إلى عدة أنظمة فرعية منفصلة بعضها عن بعض، بعد تحديده بالطبع للنماذج والبروتوكولات التي تحكم عملية الاتصال بين هذه النظم الفرعية، وقد تم تطوير برنامج لعرض هذه النماذج والبروتوكولات ولكيفية تطبيقها.

ولعل أهم ما يميز مشروع هارفست هو برنامج "الجامع Gatherer"، الذي يتولى عملية تجميع المعلومات الكشفية من مجموعات المكتبات الرقمية، مع ملاحظة أن هذا البرنامج أكثر ما تكون كفاءته عندما يتم تركيبه وتهيئته مع نظام إدارة تلك المجموعات. ويقوم كل "جامع" باشتقاق معلومات التكشيف من المجموعات وإرسالها في صيغة معيارية عن طريق بروتوكول قياسي إلى برامج تسمى "بالوسطاء أو البروكر brokers"؛ حيث يقوم الوسيط ببناء كشاف تجميعي combined index يضم المعلومات المأخوذة من كل المجموعات التي تم تكشيفها.

وتجدر الإشارة إلى أن البنية الفنية لنظام هارفست تتمتع بكفاءة عند استخدامها مع شبكة من الموارد تزيد على كفاءتها عند استخدامها مع طرق التكشيف التي تعتمد على زواحف الويب، وبالرغم من أن الفريق قد عمل على تطوير الذاكرات الفورية caches، وطرق الاستنساخ من أجل رفع كفاءة هذه النظم، فإن الفائدة الحقيقية تمثلت في البحث الجيد عن المعلومات واستكشافها. ومع أن جميع برامج "الجوامع" تقوم بنقل المعلومات في بروتوكول محدد،

يعرف بالصيغة المختصرة لتبادل الكائنات the summary Object Interchange Format (SOIF)، فإن الكيفية التي تجمع بها المعلومات يمكن تفصيلها على حسب المجموعات الفردية. وفي الوقت الذي تعمل فيه زواحف الويب على المعلومات المتاحة مجاناً open access، فإن الجوامع يمكن أن تتمتع بمزايا وصول أكثر حيث يمكن استخدامها في تكشيف المعلومات المحظورة. كما أنه يمكن تهيئتها لتناسب قواعد بيانات محددة، ولا يتطلب الأمر أن يتم تقييدها على المعلومات المتاحة على صفحات الويب أو أي صيغة محددة فقط. كذلك يمكن تضمينها قواميس أو معاجم مصطلحات لأغراض البحث في مجالات موضوعية متخصصة، ولا ريب أن جمع هذه الخصائص مجتمعة تعد مزايا كبرى لهذا النظام.

ومما تجدر الإشارة إليه أن كثيرًا من مزايا البنية الفنية لنظام هارفست تتوارى ما لم يتم تركيب "الجامع" مع مجموعات المكتبة الرقمية، ولهذا السبب فإن البنية الفنية لهارفست تعد فعالة بشكل خاص بالنسبة للمكتبات الرقمية الاتحادية. ففي سياق ذلك الاتحاد يمكن أن تقوم أي مكتبة بتشغيل الجامع الخاص بها، والذي يقوم بنقل المعلومات الكشفية إلى "الوسطاء (البروكر)" الذين يقومون بدورهم ببناء الكشافات الجامعة للمكتبة كلها، ومن ثم يتم الجمع بين مزايا عملية التكشيف المحلية من ناحية، ومزايا الكشاف المركزي من ناحية أخرى.

ما وراء البحث :

إن عملية استكشاف المعلومات تعد أكثر عمقاً من عملية البحث عن المعلومات، ولذلك فإن معظم المستفيدين يستخدمون بعض أشكال الدمج بين التصفح والبحث المنظم. وقد أشرنا في الفصل العاشر إلى المتطلبات التي ينبغي أن تتوافر للمستفيدين لكي يتمكنوا من إجراء عملية البحث عن المعلومات، وإلى صعوبة تقويم فاعلية عملية استرجاع المعلومات في إحدى جلسات البحث التفاعلية مع تواجد المستفيد في الحلقة (١) with user in the loop، ولا شك أن المكتبات الرقمية اللامركزية تشعر بحدة كل تلك المشكلات.

لقد ظل التصفح على الدوام هو الطريقة المهمة لاستكشاف المعلومات في المكتبات، ويمكن أن يكون التصفح بالبساطة التي تكون عليها عملية تصفح أرفف المكتبة للتعرف إلى الكتب التي تُجمع بعضها مع بعض. غير أن هناك وسيلة أخرى أكثر منهجية تتمثل في البدء بتصفح أحد الكتب ثم الانتقال إلى الأعمال الأخرى التي يحيل إليها هذا العمل. ومن المعروف أن معظم مقالات الدوريات وغيرها من الأعمال العلمية الأخرى تشتمل على قوائم بالإرجاعات الببليوجرافية التي تحيل إلى أعمال أخرى. ومع أن تتبع هذه الإرجاعات يعد جزءاً أساسياً من عملية البحث العلمي؛ فإنها بلا شك مهمة مرهقة، وخاصة عندما تكون المواد كائنات مادية يجب استرجاعها مكتملة في آن واحد. ولعل وجود الروابط الفائقة، يجعل تتبع الإرجاعات الببليوجرافية عملية سهلة، وهناك مقولة عامة ترى أن تتبع الروابط والإرجاعات تعد من الأمور الميسورة في سياق المكتبات الرقمية، مع أن كفاءة الفهارس

(١) لعل المؤلف يقصد دائرة التفاعل بين المستفيد ونظام البحث (المترجمان).

والكشافات عادة ما تكون أكبر في سياق المكتبات التقليدية. ومن ثم فمن المحتمل أن يكون التصفح أكثر أهمية - بشكل نسبي - في سياق المكتبات الرقمية.

وإذا تتبع المستفيدون توليفة تجمع بين التصفح والبحث، مستخدمين أنواعاً متفرقة من المصادر ومحركات البحث، فما هي إذن درجة الثقة التي سيولونها للنتائج التي يتحصلون عليها؟ لقد أشرنا في هذا الفصل إلى صعوبات المقارنة بين النتائج التي يتم الحصول عليها من خلال عمليات البحث في مجموعات مختلفة من المعلومات، وصعوبات تقرير درجة تطابق المعلومات وتكرارها في عملين تم العثور عليهما من مصادر مختلفة. ويواجه المستفيدون الجادون من المكتبات الرقمية مشكلة بقدر ما هي دقيقة فإنها تعد أكثر خطورة، وهي أنه من الصعوبة غالباً أن نعرف مدى شمولية البحث الذي يتم تنفيذه؛ فالمستفيد الذي يقوم بالبحث في أحد مراصد البيانات المركزية، كنظام الميدلاين Medline الذي ترعاه المكتبة الطبية، يمكن أن يكون على درجة كبيرة من الثقة في أن البحث أجري لكل تسجيلة من تسجيلات النظام. قارن ذلك مع بحث موزع يجري في عدد كبير من مجموعات البيانات، وتساءل: ما فرصة ضياع معلومات هامة؟ هل لأن إحدى مجموعة البيانات تختفي في مجموعات أخرى عند تقديم المعلومات الكشفية، أم لأن إحدى مجموعات البيانات تخفق في الرد على التساؤلات البحثية ؟

إن عملية البحث الموزع تلخص الوضع الحالي للمكتبات الرقمية، فمن إحدى وجهات النظر، يمكن القول بأن أي أسلوب فني لا يخلو من نقاط ضعف خطيرة، وخاصة أنه لم تظهر بعد المعايير الفنية الكاملة المتفق عليها، كما أن فهم حاجات المستفيدين لا يزال في مراحله التمهيدية، يضاف إلى ذلك أن الصعوبات التنظيمية لا تزال متفاقمة. وعلى الرغم من ذلك، وفي الوقت نفسه، هناك كميات هائلة من المعلومات متاحة على الإنترنت، كما أن برامج بحث الويب متاحة مجاناً، بل إن الاتحادات والخدمات التجارية تتسع بشكل متزايد. وأخيراً يمكن القول إنه عن طريق الجمع الذكي بين عمليتي البحث والتصفح، يستطيع المستفيدون الشغوفون أن يصلوا إلى المعلومات التي يبحثون عنها.