Discovery of Services Based on WSDL Tag Level Combination of Distance Measures

*Benjamin Kanagwa and Agnes F. N. Lumaala**
School of Computing and Informatics Technology, Makerere Univesity

ABSTRACT

During discovery of services, a service request is compared with available services. A similarity measure is one of the techniques which can be used to quantify the comparison and to rank existing services according to their suitability to satisfy a given request. Similarity measures have varying strengths and are applied in isolation. We argue that we can have better ranking if we apply the strongest similarity measure to a specific element of the service description. That is, we can employ a number of similarity measures to a specific service. In this paper we experiment with application of a combination of similarity measures to specific elements in a WSDL document. We then aggregate the similarity values into a single metric that is used to give the overall ranking of the services. Also, in our approach, the contribution of each element in the WSDL document is weighted according to the significance of the element in describing the desired service.

Keywords: service matching, similarity measures, service discovery, web services.

IJCIR Reference Format:
Kanagwa, Benjamin and Lumaala, Agnes F.N. Discovery of Services Based on WSDL Tag Level Combination of Distance Measures. International Journal of Computing and ICT Research, Vol. 5, Special Issue, pp 17-24. http://ijcir.org/specialissue2011/article3.pdf

1. INTRODUCTION

Agility, interoperability, incremental integration, re-usability and flexibility have created enormous research and industrial interest in the domain of web services [8, 32, 30, 27]. As a result of this research and industrial activity, a significant number of web services have been developed and made available on the World Wide Web. What is availed is the description of the web service including functional and non-functional semantics. There are various web service description languages such as WSDL [4], and semantically enriched languages such as OWL-S [23], WSMO [7], WSDL-S [28], SAWSDL [18]. The availability of these web services has created a new challenge of identifying the most suitable web service to satisfy a given request. Service discovery is the process of locating existing web services based on their functional and non-functional semantics. Techniques to discover web services offer means of matching web service consumers with web service providers [29]. Most matchmakers compare the user request to the service description by looking at common patterns.

A common characteristic of most current matchmakers is that the matching between a request and service depends on application of a single technique or combination of techniques to the whole service description document. Most matchmakers treat the web service description document as one unit and they attach equal relevance to all elements therein. We argue that some elements in the web service description document are more informative than others. Also, some elements are more structured than others. At the same time, some similarity measures are more efficient on structured elements than others. These differences in the capability of

* Benjamin Kanagwa and Agnes F. N. Lumaala. School of computing and Informatics Technology, Makerere University. P.O.Box 7062, Kampala. Emails: bkanagwa@cit.mak.ac.ug, anamulindwa@cit.mak.ac.ug

International Journal of Computing and ICT Research, ISSN 1818-1139 (Print), ISSN 1996-1065 (Online), Vol.6, Special Issue, pp 17-24. January 2012.

different techniques and relative relevance of different elements of the WSDL document can not be exploited if the description document is considered as a single unit.

Currently, for web service discovery, a given similarity measure is applied to the entire document and yet, different measures may be suitable for different segments of the documents. In this paper, we experiment by applying different combinations of algorithms to different sections of the WSDL to provide more accurate combined similarity measures. Specifically, we identify attributes within the WSDL documents and their significance in matching. We study and identify suitable algorithms to facilitate the matching of WSDL attributes to user requests. We further develop a functional platform independent implementation of a service selection and ranking of services. The choice of techniques applied is based on the accuracy of the individual algorithms. There are results of previous studies to gauge the suitability of the techniques under different situations. We also use some of the existing results to decide on the suitable algorithms for various situations. This implies that in the course of the study, we may vary the level of detail used through restricting the use of different techniques to specific parts of the document in order to improve recall or precision.

2. RELATED WORK

The subject of service discovery is well recognized in the domain of service-oriented computing. The focus of research is centered on techniques that perform accurate matching between a service request and existing services. The approaches vary both in actual technique used and the type of information used. All existing work applies logic-based, non-logic-based approaches or a combination of both. Non-logic-based approaches do not perform any inferences in service semantics. Such approaches employ several techniques including syntactic similarity matching [17], structured graph matching [25], data mining, linguistic, numerical concept distance over a given ontology [3] and generally content based retrieval. Their accuracy is dependent on the features considered during matching. Alternatively, logic based approaches rely on some logic to represent both the request and the service. Logic based approaches exploit logic based inferences on the underlying logic that represents the request or description. Representative examples of these approaches include [2, 24, 31]. Hybrid approaches such as OWL-MX [15, 16] combine logic based matching and non- logic based matching. Similarity measures [6] form part of non-logical techniques and have widely been applied in matchmaking. They easily scale from semantic matchmaking that may involve linguistic semantics and texture descriptions to functional matchmaking as used in our case.

Compared to the logic based approaches, similarity measures do not suffer from the decidability problem and can be used at different levels of matching. Similarity measures rely on traditional Information Retrieval techniques. Such techniques include *cosine* based similarity, *extended Jacquard* -based similarity, intentional *loss of information*, *Jensen-Shannon information divergence* and many others. Registries such as UDDI provide lookup services that support keyword search. Non logic-based approaches include (a) graph matching (b) data mining (c) linguistics and (d) content based retrieval among others. Klusch [15, 16] provides a hybrid approach that combines implicit service semantics and logic-based approaches. They report results superior to any of the methods used in isolation. Under implicit semantics they select the most promising similarity metrics previously ranked by [6] on the basis of their recall precision. Particularly, they mix cosine based similarity, *extended Jacquard*-based similarity, intentional *loss of information* and *Jensen-Shannon information divergence* similarity retrieval (see[15, 16] for corresponding definitions). In all the work reviewed above, there is no evidence of investigation to exploit the varying strength of similarity techniques on to specific elements in the WSDL document

3. BACKGROUND

Upon describing the service, the request is systematically matched with the WSDL document. The techniques used to compare the request to the WSDL document are based on vectors. For this reason, both request and WSDL documents may be converted into vectors using the Vector Space Model (an algebraic model).

To convert both the request and the services into vector representations, we used Term Frequency Inverse Document Frequency (TF-IDF) [17]. The TF-IDF provides pointers to which terms should be assigned lower priority due to their repeated appearance in several documents. Common terms such as “a”, “the” and “and” would be assigned lower priority and hence have insignificant impact on the final result [9].

3.1 Similarity Coefficients

In this section, we highlight specific distance measures that were selected based on the current usage in different areas of information retrieval. The *Jaccard similarity coefficient* [11] is a metric used for comparing the similarity and diversity of sample sets. It achieves this through comparing the intersection with the union of the sample sets, otherwise illustrated as: $J(A,B)=\frac{|A\cap B|}{|A\cup B|}$. Consequently, are Jaccard coefficient of 1 implies perfect match while 0 implies no match.

The *Cosine Similarity* [13] is described as a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them. This may be illustrated with two vectors of attributes, A and B, where the cosine similarity, (θ), is represented using a dot product and magnitude as

$$similarity=\cos(\theta)=\frac{A.B}{\|A\|.\|B\|}$$.

The *Hamming Distance* measures the number of differences between two vectors [14]. Alternatively, the Hamming Distance may be likened to the number of differences between two words of similar length. If q = number of variables with value 1 for the i^{th} objects and 0 for the j^{th} object, and r = number of variables with value 0 for the i^{th} objects and 1 for the j^{th} object, we have: $dij = q + r$. Consequently, a zero Hamming Distance is symbolic of a perfect match while the further away from 0 the result is, the less the similarity between two adjacent vectors or words.

Named after the Russian Scientist *Vladimir Levenshtein (*responsible for its inception), the Levenshtein or Edit Distance refers to a measure of the similarity between two adjacent strings [10]. A Levenshtein Distance of 0 symbolizes a perfect match however the further away from 0 the value, the greater the deviation between the strings. The *Damerau-Levenstein Distance* is a generalization of the Levenstein Distance. It assigns weights to the modification operation, that is, weights are assigned to the deletion, insertion or transposition operations [12].

4. EXPERIMENTAL SETUP AND RESULTS

We used a collection of 32 (See Table 1) web services by experimenting with their WSDL documents. From each WSDL document we extracted specific tag contents. We subjected the tags to a combination of distance measures. Figure 1 shows the interface used to obtain combinations of similarity techniques

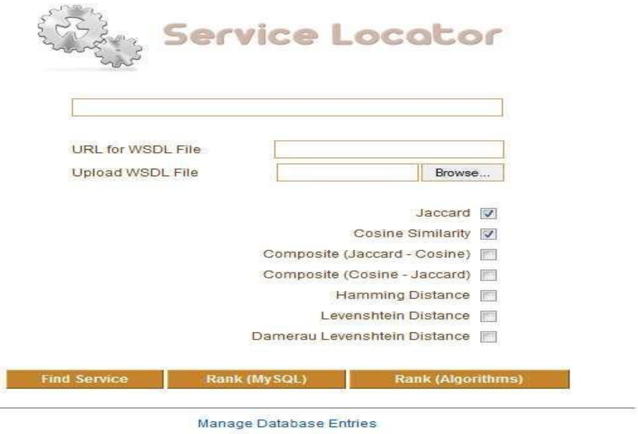

Figure 1: Interface for service locator

In order to rank multiple WSDL documents against a query or keywords, the contents of the WSDL documents were pre-loaded into a MySQL database. This was done to mitigate the challenges associated with loading more than one WSDL document at a time in order to perform a comparison. Multiple Algorithms were then integrated into the Brokerage System while ensuring that results from more than one algorithm were available for comparison. These algorithms included: i*)* Jaccard (J); *ii*) Cosine Similarity (C); *iii)* A merger of the Jaccard and Cosine Similarity(JC- This involved the application of the Jaccard to the "documentation" portion of the WSDL document, and Cosine Similarity to the other tags within the WSDL document); *iv)* A merger of the Cosine Similarity and Jaccard technique (CJ- This involved the application of the Cosine Similarity to the "documentation" portion of the WSDL document, and Jaccard to the other tags within the WSDL document); *v)*

Hamming Distance (H) ; *vi)* Levenshtein Distance (L) and *vii)* Damerau Levenshtein Distance (DL). The keywords used for the request are "city distance", parsed to the matching algorithm as an array of words "city" and "distance".

4.1Tags and References

A WSDL document is composed of different XML elements (see Table 2) that define different aspects of the service which it describes. The first three elements comprise the 'logical part' and the last two comprise the 'concrete part'. The portType defines sets of operations to be exposed. The relevance of the tags is categorized as High, Medium, and Low with 5, 3 and 1 as the respective weights.

Table 1: WSDL Document Collection

#	**WSDL**	**url**
1	postcodeservices.asmx,	http://ws.epostcode.com/uk/
2	placelookup.asmx,	http://codebump.com/services/
3	GeographicalCoord.asmx,	http://www.discoverdance.co.uk/UKCoordinates/
4	Proximity.asmx,	http://www.sircweb.cn/SircWeb/Services/Proximity/
5	addresslookup.asmx,	http://ws.cdyne.com/psaddress/
6	api.wsdl,	http://6pp.kvdb.net/services/soap/
7	location.asmx,	http://www.annotatedearth.com/AELocationService/
8	locationByZip.wsdl,	http://www.flash-db.com/services/ws/
9	zipcodelookup.asmx,	http://codebump.com/services/
10	WebService.asmx,	http://www.simplylookupadmin.co.uk/
11	RoutePlanner.asmx,	http://www.sircweb.cn/SircWeb/Services/
12	Distance.asmx,	http://voservices.net/Cosmology/ws_v1_0/
13	runnercalculator.asmx,	http://www.deeptraining.com/webservices/
14	locationByZipService.wsdl,	http://www.flash-db.com/services/ws/
15	Nearest.asmx,	http://www.allies-computing.co.uk/WebsoapNearest/
16	OligoSelection.wsdl,	http://projects.mi.fu-berlin.de/hobit/Oligo/
17	RoadDistance.wsdl,	http://rfservice.routefinda.com/
18	AirportInformation.wsdl,	http://www.webservicex.net/
19	AustralianPostcode.wsdl,	http://www.webservicex.net/
20	EMBLNucleotide Sequence.wsdl,	http://www.webservicex.net/
21	GetISBNInformation.wsdl,	http://www.webservicex.net/
22	HCPCS.wsdl,	http://www.webservicex.net/
23	ICD9Drug.wsdl,	http://www.webservicex.net/

25	ICD-9-CM.wsdl,	http://www.webservicex.net/
26	ICD-10-CM.wsdl,	http://www.webservicex.net/
27	MediCareSupplier.wsdl,	http://www.webservicex.net/
28	UKLocation.wsdl,	http://www.webservicex.net/
29	USAddressverification.wsdl,	http://www.webservicex.net/
30	USAZipcode Information.wsdl,	http://www.webservicex.net/
31	BibCode.Query.wsdl,	http://adsabs.harvard.edu/
32	Catalog.Search. Services.wsdl,	http://irsa.ipac.caltech.edu/

Table 2: A WSDL document's primary attributes

#	**Tag**	**Relevance**	**Justification**
1	<portType>	Medium	Assigned Medium priority because most portType attributes contain technical descriptions for port functionality e.g. “BuiltInTypeSoapPort”. Such descriptions provide little relevance to users during semantic matching.
2	<message>	Medium	Assigned Medium priority because most message attributes refer to technical descriptions for messages within the WSDL document e.g. “BasicComplexType”. Such descriptions provide little relevance to users during semantic matching.
3	<documentation >	High	Assigned High priority because documentation tags contain the most descriptive attributes of a WSDL document. These descriptions provide a lot of relevance to users during semantic matching.
4	<targetNamespa ce>	Low	Assigned Low priority URLs often contain few descriptive attributes of a WSDL document. As a result, URLs provide little relevance to users during semantic matching.

4.2 Ranking Similarity Combinations

To gauge the superiority of the different combinations of similarity measures, we applied a correlation coefficient. To provide basis for comparison, we first ranked the services manually based on our judgment of the best service for the request. Then we used Pearson’s Correlation Coefficient to show how well a given combination of distance measures fared against the base ranking. In table 4, the column labeled HB represents the base ranking against which all other rankings are compared. The Pearson’s Correlation Co-efficient measures the strength of association between two variables or range of variables [26]. The correlation coefficients can be interpreted as follows: -1.0 to -0.7 strong negative association; -0.7 to -0.3 weak negative association; -0.3 to +0.3 little or no association; +0.3 to +0.7 weak positive association; +0.7 to +1.0 strong positive association.

Table 4: Interpretation of results

#	**WSDL**	**HB**	**J**	**C**	**JC**	**CJ**	**H**	**L**	**DL**
1	Catalog.wsdl	1	30	17	17	29	27	22	30

2	BibCode.Query.wsdl	2	3	13	19	3	28	19	28
3	USAZipcodeInformation.wsdl	3	2	7	13	1	26	9	27
4	USAddressverification.wsdl	4	31	32	32	32	32	1	32
5	UKLocation.wsdl	5	32	31	26	31	31	2	31
6	MediCareSupplier.wsdl	6	29	20	31	19	29	3	29
7	ICD-10-CM.wsdl	7	8	10	16	4	25	24	24
8	ICD-9-CM.wsdl	8	1	8	10	2	20	29	25
9	ICD9ToICD10.wsdl	9	10	6	5	10	8	5	21
10	ICD9Drug.wsdl	10	5	4	7	6	5	8	20
11	HCPCS.wsdl	11	6	1	3	5	3	21	18
12	GetISBNInformation.wsdl	12	10	3	6	9	7	28	19
13	NucleotideSequence.wsdl	13	18	2	1	14	1	32	16
14	AustralianPostcode.wsdl	14	7	5	2	12	2	27	15
15	AirportInformation.wsdl	15	24	25	22	26	16	12	6
16	RoadDistance.wsdl	16	17	26	24	21	18	6	3
17	OligoSelection.wsdl	17	19	28	28	23	19	10	2
18	locationByZip.wsdl	17	10	9	4	15	4	26	14
19	Nearest.asmx	17	14	12	8	16	6	30	13
20	runnercalculator.asmx	17	16	15	11	17	9	25	12
21	Distance.asmx	17	20	18	12	18	11	17	11
22	RoutePlanner.asmx	17	27	24	23	27	15	13	7
23	WebService.asmx	17	25	23	20	24	14	14	8
24	locationByZip.wsdl	17	22	21	14	20	12	16	10
25	zipcodelookup.asmx	17	23	22	18	22	13	15	9
26	location.asmx	17	21	30	30	25	22	20	1

27	api.wsdl	17	9	16	25	7	30	7	26
28	addresslookup.asmx	17	15	11	9	11	10	31	17
29	Proximity.asmx	17	13	19	21	13	24	23	22
30	placelookup.asmx	17	26	27	27	28	17	11	5
31	GeographicalCoord.asmx	17	28	29	29	30	21	4	4
32	postcodeservices.asmx	17	4	14	15	8	23	18	23
	PEARSON'S CORRELATION COEFFICIENTS	**-**	**0.078**	**0.191**	**-0.018**	**0.192**	**-0.456**	**0.195**	**-0.784**

5. CONCLUSION

The wide adoption of WSDL and other text based description frameworks for describing services makes it incumbent that new ways of extracting relevant details are designed. In this paper, we provide a new approach to discovery of services WSDL by applying distance measures to tags other than the entire document. This approach allows combining distance measures for optimal results.

Our study further embraces the hybrid scheme by attempting multiple combinations of techniques in a bid to improve recall and precision. In conclusion, despite the known limitations of syntactic matching, we believe it offers an effective benchmark to gauge the relevance of platform independent service brokerage. Initial results with proposed approach showed that it is an effective mechanism to obtain a reduced set of syntactically relevant services with pre-defined importance and that are trustworthy.

REFERENCES

ALGERGAWY A., NAYAK R., SIEGMUND N., K¨OPPEN V., AND SAAKE G. 2010. Combining schema and level-based matching for web service discovery. In P*roceedings of the 10th International Conference of Web Engineerin*g, Vienna, Austria pages 114–128, July 5-9. Springer-Verlag, Berlin.

BENATALLAH B., HACID M., LEGER A., REY C. AND TOUMANI F. 2005. On automating web services discovery. *The VLDB Journal*, 14(1):84–96.

CHEN L. AND XU L. 2010. A framework for web service discovery based on ontology similarity. In *Proceedings of the 2010 Fifth IEEE International Symposium on Service Oriented System Engineering, SOSE '10*, pages 197–201, Washington, DC, USA. IEEE Computer Society.

CHINNICI R., MOREAU J., RYMAN A., AND WEERAWARANA S. 2007. Web services description language (wsdl) version 2.0 part 1: Core language. http://www.w3.org/TR/wsdl20/,. W3C Recommendation 26 June 2007.

CIMPIAN E., MOCAN A., AND STOLLBERG M. 2006. Mediation enabled semantic web services usage. In *proceedings of the 1st Asian Semantic Web Conference (ASWC2006)*, September 3-7, Beijing, China 2006, pages 459-473.

COHEN W. W., RAVIKUMAR P., AND FIENBERG S. E. 2003. A comparison of string metrics for matching names and records. In *proceedings of KDD Workshop on Data Cleaning*, August 7-12, Washington DC, USA, pp. 73-78.

ROMAN D., LAUSEN H. AND KELLER U.(Eds) 2006: Web Service Modeling Ontology (WSMO), WSMO Working Group Draft. http://cms-wg.sti2.org/TR/d1/v1.0/ Accessed November 2010.

ERIN C.. Web services: Benefits, challenges, and a unique, visual development solution. www.altova.com/whitepapers/webservices.pdf .

GARCIA E. 2006. Description, advantages and limitations of the classic vector space model, http://www.miislita.com/.

GILLELAND M. 2004. Levenshtein distance, in three flavors. http://www.merriampark.com/ld.htm.

HAVELIWALA T. H., Gionis A., Klein D., and Indyk P. 2002. Evaluating strategies for similaritysearch on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web,* pages 432–442, New York, NY, USA,ACM.

HONEGGER F. 2001. Using the damerau-levenshtein-metric for string comparsion, a delphi implementation. www.delphi3000.com/articles/delphi3000_2240.htm

CHAPMAN J.L. 2000. Tanimoto coefficient, methods for ranking.

KARDI T. 2006. Similarity measurement. http://people.revoledu.com/kardi/tutorial/similarity/.

KLUSCH M. AND FRIES B. 2007. Hybrid owl-s service retrieval with owls-mx: Benefits and pitfalls. In *Proceedings of the SMRR 2007 Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web (SMRR) co-located with ISWC 2007 + ASWC 2007,* Busan, South Korea, November 11.

KLUSCH M., KAPAHNKE P., AND FRIES B. 2008. Hybrid semantic web service retrieval: A case study with owls-mx. In *Second IEEE International Conference on Semantic Computing* Santa Clara, CA, USA - August 4-7, pages 323–330.

KOKASH N. 2006. A comparison of web service interface similarity measures. In *Proceedings of the 2006 conference on STAIRS 2006: Proceedings of the Third Starting AI Researchers' Symposium IOS* Press Amsterdam, The Netherlands.

KOPECKÝ J., VITVAR T., BOURNEZ C., AND FARRELL J. 2007. SAWSDL: Semantic annotations for wsdl and xml schema. *Internet Computing, IEEE*, 11(6):60–67.

LIU M., SHEN W., HAO Q., YAN J., AND GAO Q. 2008. A multi-level matching framework for semantic web services in collaborative design. In *Proceedings of the Second International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web,* Karlsruhe, Germany, October 27, , pages 392–398.

MAEDCHE A. AND STAAB S. 2001. Ontology learning for the semantic web. *IEEE IntelligentSystems,* 16(2):72–79.

MEDITSKOS G. AND BASSILIADES N. 2007. Object-oriented similarity measures for semantic web service matchmaking. In *proceedings of Fifth IEEE European Conference on Web Services ECOWS,* pages 57–66.

SRINIVASAN N., PAOLUCCI M., SYCARA K. 2004. Adding OWL-S to UDDI, Implementation and Throughput, *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition*, USA.

OWL Group wl-s: Semanticmarkup for web services,2006.(accessed Sept. 28, 2008), http://www.ai.sri.com/daml/services/owl-s/1.2/overview/.

PAOLUCCI M., KAWAMURA T., PAYNE T.R., AND SYCARA K.P. 2002. Semantic matching of web services capabilities. In *ISWC '02 Proceedings of the First International Semantic Web Conference on The Semantic Web Springer-Verlag London, UK*, pages 333–347.

RINGE S. AND GADGE J. 2010. Semantic web services matchmaking using bipartite graph matching with ranking. In *Proceedings of the International Conference and Workshop on Emerging Trends in Technology,ICWET '10,* pages 487–492, New York, NY, USA, ACM.

RODGERS J.L.AND NICEWANDER W.A. 1998. Thirteen ways to look at the correlation coefficient, *the American Statistician*, 42(1):5966, february.

STAAB S., AALST W., BENJAMINS V.R., SHETH A., MILLER J.A., BUSSLER C., MAEDCHE A., FENSEL D., AND GANNON D. 2003. Web services: Been there, done that? *IEEE Intelligent Systems*, 18(1):72–85.

STEVE H. Wsdl-s submission request to w3c, http://www.w3.org/submission/2005/10/, accessed: Aug,09. 2005.

TSALGATIDOU A., PILIOURA T. 2002. Recommended Boualem Benatallah, and Fabio Casati. An overview of standards and related technology in web services. *Distributed and Parallel Databases* 12(2-3), Kuwer Academic Publishers Hingham, MA, USA .

WEERAWARANA S., CURBERA F., LEYMANN F., STOREY T., AND FERGUSON D.F. 2005. W*ebServices Platform Architecture : SOAP, WSDL, WS-Policy, WS-Addressing, WS-BPEL, WS-Reliable Messaging, and More.* Prentice Hall PTR, March.

WU D., PARSIA B., SIRIN E., HENDLER J., AND NAU D. 2003. Automating daml-s web services composition using shop2. In *Proceedings of 2nd International Semantic Web Conference (ISWC2003)*, 20-23 October 2003, Sanibel Island, Florida, USA .

Acknowledgement
We acknowledge all the efforts by Samuel Otim in realising the implementation of this work. Thank you for moving the work forward.