PRESUPUESTO:

Dirección de proyecto y desarrollo de diversos módulos para el proyecto **BUAALA.TV "la tv inteligente, interactiva e internet"**

INDICE

1. Objeto del Presupuesto

El cliente Agrupación Empresarial Innovadora Knowdle Consortium Group, (en adelante KCG) con C.I.F.: G-33988304, con sede en Centro Municipal de Empresas de Gijón, Avenida de Argentina 132, Gijón, 33313-Asturias y en su nombre y representación, D. Carlos Vega García con DNI 071510177-A, actuando en calidad de Gerente, ha solicitado a ITelligent Information Technologies SL las siguientes tareas que son objeto de este presupuesto:

- Dirección y oficina técnica del proyecto BUAALA.TV "la tv inteligente, interactiva e internet"

- Desarrollo de una serie de módulos que permitan extraer datos de las fuentes sobre las que actuará el proyecto y convertir estos datos no estructurados en complejos objetos de información que puedan ser "consumidos" en otras etapas del proyecto.

- Desarrollo de unos módulos que permitan ofrecer como test-case unas novedosas KAPPS, que ofrezcan una información altamente accionable (en forma de acciones o recomendaciones) a los usuarios.

2. Descripción Presupuesto

2.1. Dirección de Proyecto y Oficina Técnica

La Oficina de Direccion de Proyectos ofrecerá los siguientes servicios al consorcio:

- Consolidación de la Información.
- Gestión de la Metodología.
- Formación y Capacitación.
- AuditoriadeProyectos.
- Gestión de Tesorería.
- Valoración de Indicadores.
- Gestión de la Organización.

La dirección de proyecto se realizará atendiendo a los requerimientos metodologicos y normativos que se fijen para el proyecto (CMMI, ISO-15504, recomendaciones PMI, ...).

2.2. Módulos de Procesamiento Datos no Estructurados

Consistirá en una "pipeline", que comenzará con el acceso y extracción de los datos de aquellas fuentes de interés (ej. foros, blogs, prensa digital, ...), para posteriormente someterlos a distintos procesamientos, con el objeto de convertir los datos no estructurados en objetos complejos de información que puedan ser "consumidos" en etapas posteriores del proyecto.

Los distintos módulos que compondrán la "pipeline" se sitúan en el actual estado del arte en el área de Procesamiento del Lenguaje Natural (PLN), siendo el objetivo último la obtención de información con el nivel de calidad y estructura que el proyecto requiere con una mínima intervención manual. A continuación se describen cada uno de estos módulos:

M1-Recopilación Automática de Datos: La extracción de datos de paginas web se realiza utilizando wrappers, que son componentes designados para acceder a documentos HTML y recolectar contenidos relevantes de los mismos [Lae02]. Inicialmente los wrappers estaban diseñados para facilitar al programador la escrituras de reglas de extracción, mientras que el estado actual del arte permite la utilización de técnicas de aprendizaje automático para la generalización de reglas de extracción (ej. wrapper induction systems). La principal innovación del sistema propuesto en este proyecto será su habilidad para extraer datos útiles de diferentes fuentes de información, con distintos esquemas y su capacidad para lidiar con cambios en las paginas webs, para ello el sistema chequeará la aparición de nuevos contenidos y combinará estas novedades con los datos previamente obtenidos. Otra importante innovación del sistema será la combinación de los tres enfoques básicos [Kay06] para la construcción de wrappers (manual, inducción supervisada y no supervisada), incorporando un mecanismo de aprendizaje que permitirá al sistema decidir cuándo reducir el grado de automatismo (de

automático a semiautomática o manual), manteniendo un equilibrio óptimo entre eficacia y esfuerzo manual. El sistema propuesto permitirá el rápido despliegue de wrappers para las fuentes de datos gestionadas por el proyecto, y la extracción diaria de los datos de esas fuentes. El aspecto clave es la reducción drástica del esfuerzo manual mediante la detección de los casos en que realmente se necesita, y proporcionar al usuario información importante (como dónde y por qué la intervención manual es necesario).

[Kay06] M. Kayed, M.R. Girgis, and K.F. Shaalan, "A Survey of Web Information Extraction Systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, Oct. 2006, pp. 1411-1428.

[Lae02] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira, "A brief survey of web data extraction tools," *ACM Sigmod Record*, vol. 31, 2002, p. 84–93.

M2-Clasificación Automática de Textos: Este modulo permite asignar una categoría (de un conjunto predefinido de ellas) a un texto en lenguaje natural. La forma más común de resolver esta tarea es utilizando algoritmos de aprendizaje sobre una colección (corpus) significativo de documentos de cada categoría. El sistema que se propone en este presupuesto introduce varias innovaciones a la aproximación habitual, principalmente estas innovaciones están orientadas a reducir el esfuerzo manual requerido para la creación de los clasificadores. Por ejemplo, permitiendo a los usuarios la creación de un "golden-standard" de documentos pre-clasificados de acuerdo con una taxonomía definida por el cliente, utilizando algoritmos de expansión [Mit98] que permitan generar nuevas "features" para una determinada categoría, o utilizando métodos para extraer conocimiento utilizable [Esu07] de grandes colecciones de textos no anotados (en vez de crear una colección de documentos anotados para cada nueva categoría incluida en el sistema). Otra importante innovación es la integración en el sistema de los usuarios mediante un proceso de creación de conocimiento en colaboración, por ejemplo permitiendo a los usuarios definir sus propias categorías (folksonomies [Pet09]). El resultado será un sistema de clasificación capaz de reducir al mínimo la intervención humana en la creación de nuevos clasificadores y que ofrecerá a los usuarios funcionalidades avanzadas como: agrupar contenidos dinámicamente dependiendo en los resultados del clasificador, creación de colecciones personalizadas de documentos [Ado05]., enriquecimiento de contenidos añadiendo nuevos conceptos y/o categorías, etc...

[Ado05] G. Adomavicius, A. Tuzhilin. *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.* IEEE Transactions on Knowledge and Data Engineering. 2005.

[Esu08] A. Esuli, F. Sebastiani. *PageRanking WordNet Synsets: An Application to Opinion Mining.* In Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics . 2007.

[Mit98] M. Mitra, A. Singhal, C. Buckley. *Improving automatic query expansion.* Proceedings of the 21st annual international ACM SIGIR conference on Research and development in Information Retrieval. 1998.

[Pet09] I. Peters. *Folksonomies: indexing and retrieval in Web 2.0.* De Gruyter/Saur, Berlin. 2009.

M3-Extracción Automática de Información: Este modulo permite la extracción y desambiguación [Cuc07] de entidades nominales (EN) presente en documentos (ej. el nombre de una persona, nombre de una empresa, una dirección, etc...), la extracción de relaciones (ER) entre entidades nominales [Ban08] y el descubrimiento de nueva información sobre las entidades nominales. Actualmente el estado del arte utiliza algoritmos de aprendizaje supervisados lo que requiere el esfuerzo manual de anotar datos que sirvan como ejemplos para el aprendizaje. En este presupuesto además del uso de algoritmos de aprendizaje supervisado se va a investigar algoritmos no supervisados que permitan crear un ranking que ayude a reducir el esfuerzo manual. El resultado serán documentos anotados con entidades nominales [Zha10], que habrán sido previamente desambiguadas, utilizando links con formato URIs únicos. Cada entidad nominal será incorporada a una base de datos, permitiéndose la detección de relaciones entre entidades nominales, a partir de las cuales se establecerá una red de documentos relacionados. Además se exploraran técnicas no supervisadas para la detección de nueva información (atributos) de las entidades nominales, mediante la creación automática de patrones [Bri98] (ej. X es el director comercial de la empresa Y), que permitan utilizar los recursos web disponibles (ej. Google) para inferir nuevos atributos y relaciones de las entidades localizadas.

[Cuc07] S Cucerzan. 2007. Large-Scale named entity disambiguation based on Wikipedia data. Proc. of EMNLP-CoNLL '07.

[Zha10] W. Zhang, J. Su, C. L. Tan, and W. T. Wang. 2010. Entity linking leveraging automatically generated annotation. Proc. of Coling 2010.

[Bri98] S. Brin. 1998. Extracting Patterns and Relations from the World Wide Web. WebDBWorkshop at
EDBT'98.

[Ban08] M. Banko und O. Etzioni. 2008. The Tradeoffs Between Open and Traditional Relation Extraction. Proc. of ACL-08: HLT.

M4-Minería de Opinión: Se trata de un área de investigación emergente que se ocupa de las opiniones, emociones y la subjetividad en los textos. Las investigaciones más recientes (estado del arte) se centran principalmente en sistemas de minerías de opinión independientes del dominio. La principal innovación del sistema propuesto es la definición de un "kernel" de extracción de opinión que se adapta fácilmente a diferentes dominios a través de un conjunto de recursos específicos de cada dominio. La inducción

de estos recursos se realizará automáticamente a partir de una pequeña colección de documentos comentados. El sistema permitirá estructurar la información subjetiva procedente de foros, blogs, noticias, etc...entre ellas:

- Detectar la subjetividad en textos [Jan04]
- Clasificar dicha subjetividad, según la opinión sea positiva o negativa[Tur03], o según una perspectiva ideológica [Mul08]
- Extracción de la representación estructurada de las opinions individuales [Cru10] (ej. el objeto de la opinión, la polaridad de la opinión, titulares de la opinión, ...)
- Detección del estado de animo y emociones en los textos subjetivos [Ove05]

[Jan04] Janyce M. Wiebe et al., "Learning Subjective Language", Computational Linguistics, 2004

[Tur03] Peter D. Turney and Michael L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association", ACM Transactions on Information Systems, 2003

[Mul08] Tony Mullen and Robert Malouf, "Taking sides: User classification for informal online political discourse", Internet Research, 2008

[Cru10] Fermín L. Cruz et al., "A Knowledge-Rich Approach to Feature-Based Opinion Extraction from Product Reviews", 2nd International Workshop on Search and Mining User-Generated Contents (ACM), 2010

[Ove05] Cecilia Ovesdotter Alm et al., "Emotions from text: machine learning for text-based emotion prediction", Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, 2005

2. 3. Módulos de Información Accionable

El proyecto, BUAALA.TV "la tv inteligente, interactiva e internet" plantea la creación de KAPPS, que permitan a los usuarios acceder a aplicaciones inteligentes basadas en el conocimiento que la plataforma generará. En este presupuesto el proceso de obtención de "inteligencia", se define como la forma de obtener, a partir de unos datos de partida y por procesamiento de los mismos, información elaborada o conocimiento.

En este apartado se plantea el desarrollo de varios módulos que permitan la creación de KAPPS caracterizadas por ofrecer una información accionable a los usuarios (en forma de recomendaciones). Esto es unas aplicaciones que partiendo de datos no estructurados, propongan al usuario acciones óptimas en formas de recomendación. Estos módulos servirán como test-case para el futuro desarrollo de nuevas KAPPS basadas en los mismos conceptos. Estos módulos consistirán en:

M5-Modelos Probabilisticos: Desarrollar modelos probabilisticos avanzados basados en modelos jerárquico bayesianos [Bla91], [Ros03], que permiten su aplicación a distintos problemas de gran interés comercial (ej. estimación de la elasticidad de la demanda).

[Bla91] Blattberg, RC, George, E. "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations". *Journal of the American Statistical Association*. 1991; 86: 304-315.

[Ros03] Rossi, PE, Allenby, GM. "Bayesian Statistics and Marketing". *Marketing Science*. 2003; 22 (3): 304-328.

M6-Modelos de Optimización: Desarrollar técnicas de optimización mediante programación no lineal que permitan determinar el valor de los parámetros que maximicen la función de utilidad del usuario (ej. el margen de contribución).

M7-Modelos de Soporte: Desarrollar los algoritmos y la plataforma (principalmente por paralelización) que permita resolver los problemas planteados en una ventana de tiempo adecuada.

3. Presupuesto

Concepto	Unidad	Coste Unitario	Nº Unidades	Total
Direccion de Proyecto y Oficina Tecnica				
Direccion				93.750,00 €
Total->				**93.750,00 €**
M1-Recopilación Automática de Datos				
Direccion	horas	45,00 €	100	4.500,00 €
Analista	horas	39,00 €	231	9.009,00 €
Programador	horas	30,00 €	825	24.750,00 €
Colaboracion (Univ. Sevilla)	Euros			6.752,00 €
Total->				**45.011,00 €**
M2-Clasificación Automática de Textos				
Direccion	horas	45,00 €	86	3.870,00 €
Analista	horas	39,00 €	202	7.878,00 €
Programador	horas	30,00 €	722	21.660,00 €
Colaboracion (Univ. Sevilla)	Euros			5.908,00 €
Total->				**39.316,00 €**
M3-Extracción Automática de Información				
Direccion	horas	45,00 €	94	4.230,00 €
Analista	horas	39,00 €	216	8.424,00 €
Programador	horas	30,00 €	773	23.190,00 €
Colaboracion (Univ. Sevilla)	Euros			6.330,00 €
Total->				**42.174,00 €**
M4-Mineria de Opinión				
Direccion	horas	45,00 €	113	5.085,00 €
Analista	horas	39,00 €	260	10.140,00 €
Programador	horas	30,00 €	928	27.840,00 €
Colaboracion (Univ. Sevilla)	Euros			7.596,00 €
Total->				**50.661,00 €**
M5-Modelos Probabilisticos				
Direccion	horas	45,00 €	89	4.005,00 €
Analista	horas	39,00 €	205	7.995,00 €
Programador	horas	30,00 €	735	22.050,00 €
Colaboracion (Univ. Cadiz)	Euros			5.993,00 €
Total->				**40.043,00 €**
M6-Modelos de Optimización				
Direccion	horas	45,00 €	81	3.645,00 €
Analista	horas	39,00 €	188	7.332,00 €
Programador	horas	30,00 €	671	20.130,00 €
Colaboracion (Univ. Cadiz)	Euros			5.485,00 €
Total->				**36.592,00 €**
M7-Modelos de Soporte				
Direccion	horas	45,00 €	31	1.395,00 €
Analista	horas	39,00 €	106	4.134,00 €
Programador	horas	30,00 €	414	12.420,00 €
Colaboracion (Univ. Cadiz)	Euros			5.500,00 €
Infraestructura				4.135,00 €
Total->				**26.189,00 €**

Gran Total ->	373.736,00 €
IVA (18%) ->	67.272,48 €
Gran Total con Iva ->	441.008,48 €

Este presupuesto asciende a un total de trescientos setenta y tres mil setecientos treinta y seis euros (373.736 €) Iva no incluido.

4. Forma de Pago y Plazo

Se facturará el 25% del presupuesto al inicio del proyecto, la parte restante (75%) se realizará de la siguiente forma:

- Dirección de Proyecto: Se facturará a la recepción del proyecto.
- Módulos: Se facturarán a la entrega de cada uno de ellos según calendario a pactar con el cliente.

Otros:

- Las facturas tendrán vencimiento a 60 dias desde la fecha de emisión.
- Este presupuesto sólo será válido hasta el 01-Julio-2011.

Aceptado el 6 de Mayo de 2011,

ITELLIGENT
ITelligent Information Technologies S.L.
B-11872066
www.itelligent.es

Fdo. D. Carlos Vega García
Gerente
KCG
G-33988304
Centro Municipal de Empresas
Avenida Argentina 132
33313 Gijón (Asturias)

Fdo. D. Jaime Martel R. Valdespino
Director Técnico
ITelligent Information Technologies SL
B-11872066
Parque Tecnológico TecnoBahía
Carretera Puerto-Sanlucar Km, 7,5
11500 Puerto de Santa María
(Cádiz)